The Impact of Social Groups on Variation in Infectious Disease Transmission and Control

James Daniel Munday

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy

of the

University of London

November 2020

Department of Infectious Disease Epidemiology

Faculty of Epidemiology and Population Health

LONDON SCHOOL OF HYGIENE & TROPICAL MEDICINE

Funded by the NIHR Health Protection Research Unit in Immunisation

Declaration

Statement of Own Work

I, James Munday, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, this has been indicated in the thesis. I have read and understood the School's definition of plagiarism and cheating given in the Research Degrees Handbook.

James Munday, November 2020

Abstract

Mathematical models of infectious disease are increasingly capable of capturing spatial and demographic factors in transmission. However, there has been limited evaluation of how ethnic and socioeconomic groups within a population might impact transmission, the effectiveness of interventions and inequalities in infectious disease outcomes. A large part of this challenge lies in identifying means by which information about how social groups interact can be measured and included in mechanistic models of transmission. By means of data analysis and mathematical modelling, I have investigated how social groups contribute to heterogeneity in transmission and how these factors may be captured in a model of transmission.

In the first part of this thesis I first present my evaluation of the roles of transmission and vaccination differences between social groups in creating inequalities in disease risk.

Secondly, I report my analysis of reported cases from the 2009 Influenza H1N1 outbreak to elucidate the spatial and social nature of the early stages of the outbreak.

Later I present a novel framework that I have developed for analysis of social contact of school-aged children and modelling transmission. This framework utilises national school and pupil data to simulate outbreaks over a network, explicitly accounting for school and household transmission links.

Finally, I present the application of this framework in two distinct settings: First, I assess the potential role of the school system in inequalities in influenza risk between ethnic and socio-economic groups in London. Then I investigate how connections between schools and households in the Netherlands might impact clustering of children unvaccinated against measles. Finally, I evaluate how such clustering impacts the epidemiology of measles in The Netherlands, where vaccine refusal is clearly associated to particular socio-religious communities.

I find evidence that inequalities in disease are most sensitive to differences in transmission if the pathogen has a low basic reproduction number. With higher basic reproduction numbers, inequalities are more sensitive to variation in vaccine uptake. Inequalities observed in influenza are not clearly reconciled by the school network structure, however the network may promote inequalities in incidence early in an outbreak, which may be interpreted as inequality in risk. Finally, school networks can explain the observed measles dynamics in the Netherlands well, reproducing the outbreak scale and geographical spread of cases reported in recent outbreaks.

Acknowledgements

A great deal of thanks and credit go to my supervisors, Dr Albert Jan van Hoek and Dr Katherine Atkins for inviting me to take on this PhD and providing consistent support throughout my studies. I owe deep gratitude to others at LSHTM for guidance and support Dr Sebastian Funk with whom I first explored the move towards Epidemiology and has provided mentorship an encouragement to me, Dr Petra Klepac who offered great support and inspiration at a key time in the development of the work, Prof Mark Jit for his balanced and rich insights and Prof John Edmunds for his encouragement and his key role in Analysis A.

Regarding my collaborators, I would like to thank Richard Pebody, who was most generous with his time and resources, making it possible for me to use analyse UK influenza data at Public Health England and offering insight into the response to the Influenza H1N1 outbreak in 2009. At The national institute for public health and the environment of The Netherlands (RIVM), I would like to thank Dr Don Klinkenberg, Prof. Jacco Wallinga and Dr Susan Hahne all of whom have been a great deal of support to the work on MMR in the Netherlands. Also, from the Education Executive Agency of the Netherlands (DUO) I thank Marc Meurs and Erik Fleur for their commitment an enthusiasm in providing Dutch schools data to the exact requirements I needed, certainly at no small effort.

I must also thank those who have supported me in the past, ultimately preparing me for undertaking a PhD. To Eddie Kerchinski who gave me the confidence to pursue mathematics, Dr Philippe Blondel for his kindness, encouragement and support, Mark Manzocchi whose technical knowledge and flexibility stirred my interest, and Mutahar Chalmers whose enthusiasm for new and exciting challenges inspired me to move beyond the familiar and who continues to nurture my curiosity.

I would also like to thank my family. My parents for always supporting me in where I wanted to go. To my daughter Emmy who had no choice in being born into the final stages of a PhD and has brought such joy in these last months. Finally, to my wife Katie, who has sacrificed a great deal to allow me to pursue these studies and been my firm support throughout.

Acronyms

BH	Basic Homophily
CHI	Coleman Homophily Index
CI	Confidence interval
CVT	Critical vaccination threshold
DUO	Dienst Uitvoering Onderwijs meaning “Education Executive Agency”
ECDC	European Centre for Disease Control
GLA	Greater London Authority
HAVO	Hoger Algemeen Voortgezet Onderwijs meaning “higher general continued education”
IBM	Indevidual Based Model
IMD	Index of Multiple Deprivation
LSOA	Lower Super Output Area
MMR	Mumps, Measles and Rubella vaccine
NMI	Normalised Mutual Information
ONS	Office for National Statistics
OR	Odds Ratio
PC4	four-digit postcode region
pH1N1	Pandemic Influenza A H1N1 (2009)
ROC	Receiver Operating Characteristic
RR	Relative Risk
SCH	Social Contact Hypothesis

SIR	Susceptible, Infected, Recovered (model)
SR	Spearman's Rank
UK	United Kingdon
USA	United States of America
VMBO	Voorbereidend Middelbaar Beroepsonderwijs meaning "Pre-vocational secondary education"
VWO	Voorbereidend Wetenschappelijk Onderwijs meaning "preparatory scientific education"
WCA	Women of childbearing age
WHO	World Health Organisation
wROC	Weighted Reciever Opporating Characteristic

Contents

Figures

Tables

1 Background and Introduction

1.1 Heterogeneity, transmission and control of infectious disease

Heterogeneity in transmission and control of infectious disease (between individuals, geographical location (areas/cities/countries), socio-economic class, climate and over time) can substantially impact the epidemiology of the disease affecting health outcomes and the effectiveness of healthcare interventions.

One particular area that remains broadly unclear is how social structure, created by preferential contact within certain social groups (such as religious or socio-economic groups), might create heterogeneity within particular populations. This type of heterogeneity is of interest in public health for two clear reasons:

Firstly, if the characteristics of infectious disease transmission differ between groups, inequalities in risk may emerge. Inequalities in health outcomes are moving up the agenda of public health authorities, for example monitoring and reduction of inequalities in healthcare services and interventions are now a statutory requirement in the UK[1]. Understanding how transmission dynamics may contribute to observed differences in health outcomes is an important part of resolving them[2, 3]. In addition, such differences may impact the effectiveness of control strategies, both within particular groups and in the population as a whole[4–6].

Secondly, heterogeneity in uptake of interventions between social groups will likely give rise to inequalities in infectious disease[7–9]. But can also impact its overall effectiveness. For example, if a large proportion of unvaccinated people cluster within a particular social group[10–13], there may be a higher risk of outbreaks. This diminishes prospects for control and eventual eradication of pathogens[14].

In this thesis I seek to dissect observed inequalities in infectious disease outcomes and uptake of vaccinations between social groups, specifically to improve understanding of how differences in transmission and uptake of vaccination between social groups contribute to observed and previously unexplained transmission dynamics in diverse populations. I approached this problem using a combination of analysis of disease data and novel mathematical modelling approaches aimed at accounting for transmission within and between particular social groups within a population.

1.2 Observations of heterogeneity in risk and control of infectious disease

Differences in health outcomes and mortality between social groups have been observed since the 19th century[15], and inequalities in infectious disease continue to be observed in many high-income countries[16]. A systematic review of inequalities in in infectious diseases and strategies to reduce them, identifies instances of observed inequalities every European Union member state[17]. It remains pertinent that certain groups of the population are still at disproportionately high risk of morbidity and mortality from diseases for which we have the available means to control.

The World Health Organization's (WHO) commission on the social determinants of health[18] published a report in 2008 addressing the broader issue of inequality. The development of a better understanding of the extent and cause of inequalities in health[18] was defined as a key target. The European Centre for Disease Control (ECDC) has made addressing inequalities in infectious disease a public health priority[19]. In addition, reductions in welfare spending following the recent financial crisis and the potential implications for infection prevention and control funding[19, 20] have further strengthened interest in addressing inequalities, which may be exacerbated by these changes.

Heterogeneity in risk of infectious disease

Despite general improvements in living conditions and hygiene, disparities in serious infectious disease outcomes are still observed in high-income settings[16, 17, 19]. In general, these disparities afflict the most deprived and vulnerable in society. In recent studies in the UK it has been shown that the most deprived quintile of the population are 30% more likely to be admitted to hospital with pneumonia than the most affluent quintile [21, 22]. Similar disparities are also observed in the rates of hospitalisation with acute gastroenteritis [23].

A clear example of inequalities in infectious disease outcome is the 2009 Influenza A/H1N1 (pH1N1) outbreak. Disparities were consistently observed between socio-economic and ethnic groups, in the UK and in other high-income countries.

Analysis of laboratory confirmed Influenza A H1N1 case data from the initial nine weeks (Wk 17-26) of the outbreak in London [24] shows clear disparities in the attack rates and

overall burden by socio-economic deprivation. The peak of the outbreak also appears to occur two weeks earlier in the most deprived areas despite higher prevalence in more affluent areas in the very early phases. The overall risk ratio between the most and least deprived quintiles was found to be 2.67 (95% CI 2.06–3.49).

Analysis of surveillance data from the West Midlands [25] from during the containment phase shows that 66·7% (95% CI 64·9–68·5) of laboratory confirmed cases were from the most socio-economically deprived quintile compared to just 6·2% (95% CI 5·3–7·1) from the least deprived quintile. South Asians accounted for 57·9% (95% CI 56·1–59·8) of confirmed cases, which far exceeds the proportion of the population they represent.

Analysis of H1N1 associated mortality rates in the UK [26–28], shows a higher rate in South Asians (up to Risk Ratio = 3), and in those who live in areas of high socio-economic deprivation (up to Risk Ratio = 2). This was particularly evident in the initial wave of the local epidemic. Furthermore, analysis of specific risk factors for influenza mortality shows no evidence of ethnicity [29, 30] being a significant risk factor, suggesting that differences in mortality reflect risk of infection rather than disease severity.

The disparities between ethnic groups and socio-economic status in case data could reasonably be partially attributable to variation in health seeking behaviour. However, analysis of anti-viral use does not support this [31]. Increased attack rate and an earlier epidemic peak in the London outbreak of pH1N1 are indicative of higher rates of transmission in areas of higher socio-economic deprivation[24].

Differences in reported cases of an infectious disease can be driven by a multitude of factors. These factors can either contribute to genuine inequalities in health (either differences in risk of infection or severity of disease), or differences in reporting (arising from differences in healthcare practice [32] and variation in health-seeking behaviour). Understanding the drivers of observed disparities is vital for development of practicable and effective mitigations against inequalities in future.

Heterogeneity in control of infectious diseases

Vaccination is the single most effective means of preventing and controlling infectious disease[33]. Whilst most health-care interventions act to improve the health of the individual, the dynamic properties of infection allow vaccination to protect unvaccinated individuals in the population indirectly through a mechanism known as herd immunity [34]. The herd immunity phenomenon means that vaccinating a large enough sub-set of the population can interrupt transmission effectively. The proportion requiring vaccination is known as the critical vaccination threshold, which has a theoretical value of $1-\frac{1}{R_0}$ [35], where R_0 is the basic reproduction number of the infection in the population affected. Indirect effects are typically of benefit to the population as a whole [36]. However, this simple model of herd immunity is conditional on homogenous transmission in the population, and uniformly distributed vaccination uptake. Where heterogeneity in transmission exists due to population structure this simple assumption breaks down. Moreover, studies of the dynamics of vaccination in realistic populations have shown significant differences in behaviour depending on demographic and social differences between contexts [5, 6]. Moreover, there are many examples where clustering of unvaccinated individuals leads to coverage lower than the critical vaccine threshold at

a local level[13]. This may not be adequately represented in regional or national estimates of vaccination coverage[37]. If these complexities are ignored vaccination can have unintended effects and in some cases has been observed to increase morbidity [9, 38].

Vaccination delivery requires complex logistical programs; there are multiple ways that heterogeneity can arise in this process[39–41]. A key factor in implementation is access and cooperation from members of the population. Difficulty in accessing health care, and negative sentiment towards certain interventions are key drivers of sub-optimal uptake. In some cases, social factors such as socio-economic status, ethnicity and religion are predictors of low uptake. Disparities in vaccine uptake between socio-economic, ethnic and religious groups have been identified in multiple settings [13].

Orthodox Jewish populations have been observed to have much lower uptake of the Measles, Mumps and Rubella vaccine (MMR) than that of the rest of the local population in many settings [42]. Notably, this lead to a measles outbreak in 2019 in the New York population [43] and multiple outbreaks in other countries in recent years [11, 12, 44, 45]. Another population with particularly low uptake of MMR are the Anthroposophic [46] and Orthodox Protestant (Dutch Reformed Church) populations [47] in the Netherlands. These populations have also experienced outbreaks in recent years [48, 49] despite consistently high national MMR uptake for the past 20 years [50].

Other factors that may contribute to clustering of unvaccinated children are less explicit. Bensal et al. showed that in the states there was a correlation between income and other social factors and low vaccination uptake, as well as significant geographical clustering in the USA [51].

Combined heterogeneities – a complex problem

Heterogeneities in transmission and uptake of vaccination combine to form a complex epidemiological problem. Since the overall impact of vaccination in itself is affected by the local transmission properties of the population. This means that the drivers of heterogeneity in disease are not easily understood or studied. By capturing the mechanism of transmission explicitly, mathematical models are well placed for studying this problem[2].

1.3 Mathematical modelling to explore infectious disease dynamics in heterogeneous populations

A brief introduction to the principles of mathematical modelling

Mathematical models have become a popular method for studying the dynamics of infectious disease. They have been used to develop much of the current understanding of infectious disease epidemiology [52]. A popular approach to modelling infectious disease transmission, is to use compartmental models. In these models, parts of the population exist in one of a number of disease states at any one time. An example of this was first proposed by Ross and Hudson [53] and popularised by Kermack and McKendrick in a series of papers in the 1920s and 30s [54–56]. This work built on Bernoulli's state space model, the most important extension was to introduce dependency between rate of transmission and incidence (1766) [57].

The basic compartmental model framework is known as the "Susceptible, Infected, Recovered" (SIR) model (Figure 1.1). The model in its most familiar form is written as a system of ordinary differential equations. In this model parts of the population are assumed to be in either the Susceptible, Infected or Recovered state. The infected part of the population can infect the susceptible part of the population at a rate known as the force of infection (λ), hence the population moves from the susceptible state to the infected state at this rate. The force of infection is proportional to the section of the proportion population that is in the infected state. In turn the part of the population in the infected state recovers becoming neither infectious nor susceptible in the recovered state. The population moves from the infected state to the recovered state at a constant rate (ρ).

$$\dot{S} = -\lambda S = -\beta IS$$

$$\dot{I} = \beta IS - \rho I$$

$$\dot{R} = \rho I$$

Variations of the state-space approach have included stochasticity [58] within the model formulation allowing variation in transmission rate and recovery, and the introduction of additional states to account for variation in transmission due to the effects of intervention [59]. Although these models capture the basic dynamics of transmission, they require the assumption of homogenous transmission between all parts of the population, and contain no information about any structure within that population.

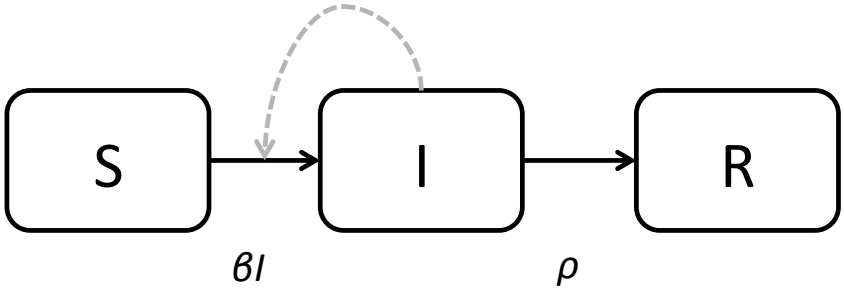


Figure 1.1 Schematic of the compartmental framework of a Susceptible (S), Infected (I) and Recovered (R) model.

Boxes show disease states indicated by letter. The population can move from state to state in the direction of the solid arrows at rates indicated by expressions below the arrow. Dashed line shows that the state affects the rate indicated.

Constructing mathematical models with heterogeneous transmission due to population structure

To study the effect of social structure on transmission with a mathematical model some approximation for the heterogeneity in transmission within and between social groups must be included in the model framework. There are a number of ways that mathematical models have incorporated heterogeneous transmission in past analyses.

A simple extension of the SIR model enables it to contain multiple compartments of each state to account for some different 'classes' of the population. A clear example is age groups[60]. In an age structured model there are separate compartments of Susceptible, Infectious and Recovered states for each age group. Force of infection in each age group is different and defined by a matrix of transmission parameters detailing rate of transmission within and between age groups (Figure 1.2). For example, separating children and adults results in a vector of force of infection expressed as:

$$\begin{bmatrix}\lambda_A\\ \lambda_C\end{bmatrix} = -\begin{bmatrix}\beta_{AA} & \beta_{AC}\\ \beta_{CA} & \beta_{CC}\end{bmatrix} \cdot \begin{bmatrix}I_A\\ I_C\end{bmatrix}$$

This approach can be extended to any number of classes, to reflect greater complexity of population structure.

Meta-population models were developed to account for transmission of a pathogen between specific groups[61–63], often geographic locations. As well as transmitting between groups, in some models, portions of the population can move between groups remaining in the same disease state. This results in infected members of the population moving into un-infected groups and causing that population to become infected.

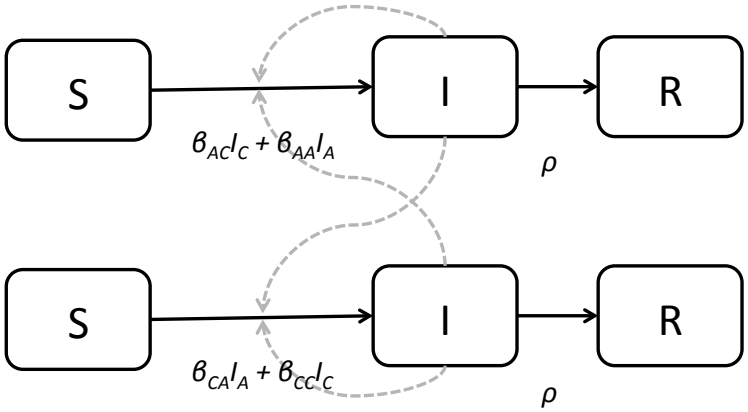


Figure 1.2 Schematic of the compartmental framework of a Susceptible (S), Infected (I) and Recovered (R) model for a simple Adult and child risk structured model.

Boxes show disease states indicated by letter. The population can move from state to state in the direction of the solid arrows at rates indicated by expressions below the arrow. Dashed line shows that the state affects the rate indicated.

Further complexity can be added by accounting for the disease status of each individual in a population separately[64, 65]. Individual-based network models (IBMs) specify a

probability of infection between hosts explicitly. IBMs have been used extensively to advance understanding of transmission dynamics on networks with particular properties from a theoretical perspective. They have also been used to create large scale 'synthetic populations' with multiple levels of population structure, such as family units, places of work and schools[66].

Parameterising population structure within models of infectious disease transmission

A key challenge with increased complexity of mathematical models is that more information is required to accurately parameterize them. In some cases, parameters can be fitted to data on infectious disease outcomes. However, increased model complexity requires a greater number of parameters to be estimated *a priori*. This parameterisation of models generally requires detailed analysis and sometimes collection of other data to provide insight into the various mechanisms within the model. Social contact surveys have become a popular method for parameterising transmission rates in and between subsets of the population.

Use of contact diaries to estimate heterogeneities in population mixing

To improve understanding of social contact structures, studies have been performed as a means to elicit information about the contact behaviour of populations. This is based on what is known as the "social contact hypothesis" (SCH)[67]. SCH is the proposition that infectious contacts, particularly in the case of respiratory infections can be assumed to be analogous to social contacts. Diary based contact surveys take a pragmatic approach by collecting self-reported contact events from participants sampled from the general

population, a method proposed and piloted by Edmunds et al in 1997[68]. For a detailed evaluation of these methods I point the reader to a recent systematic review from Thang et. al. [69]. Here I discuss contact surveys that have particular relevance to this thesis.

In 2006 Mossong et al undertook the first large scale social contact survey known as Polymod. This study was in eight European countries[70] based on this hypothesis. 7290 participants completed paper-based diaries of contact events they had made in a randomly assigned 24-hour period. Contacts were defined as close contact (conversational) or physical contact. Participants were asked to record the approximate (or exact) age of the contact, the context in which the contact was made, and the duration of the contact event. Other information regarding the relationship of the contact to the participant was also recorded. In addition, socio-demographic information was collected from each participant, such as household size and employment status.

The primary finding of the study was that age groups mix assortatively; higher contact rates exist within age groups than between age groups. Strong mixing is also shown between those of parental age and children. Schoolchildren mix most intensely and those over the age of 65 have the fewest daily contacts. The data gathered from this survey have been used to form the basis of many models where age structure is required.

In addition, qualitative observations, about how social contact networks may form can be elicited from the results. For example, contact events with contacts who are met more frequent are likely to be of longer duration, and also more likely to involve physical touch than less frequent and shorter duration events. 75 % of contacts made with strangers occur

for less than 15 minutes. Contacts were made most frequently at home and at school or work.

Following Polymod there have been 17 major diary based contact surveys in large populations that adopted a similar methodology, nine at a national level [71–79] and eight studies performed on a regional basis [80–87]. Across these studies some adjustments have been made to the survey methodology to identify characteristics of certain populations and allow deeper analysis of mixing behaviour. Differences between findings of the studies and variation within populations indicate the impact of cultural and social context in contact networks.

Some studies have included individual covariates relevant to social groups for example, employment status[80, 88], household composition[71–73, 75, 76, 81–84] and population density[81, 83]. Two studies in particular investigate social context of contacts in more depth.

A large survey of the United Kingdom[75] by Danon et. al. in 2009 investigated the nature of contact events in different social settings. The survey estimated the level of contact clustering by asking participants how many of their contacts also had contact with each other in the 7 days prior to the completing the diary. Detailed analysis of number of contacts, contact duration and clustering of contacts in different social settings allows analysis of how networks of contacts may form. In additional analysis, Danon constructed characteristic ego-networks[88] (networks centred around a specific individual) in the sampled population. The results demonstrate striking differences in the way individuals mix, based on their social and professional context. However, difficulty in connecting

ego-networks limits use of these findings in the development of large-scale models of transmission.

Danon also collected information about the distance of contact events from the participants home, a measure also undertaken by Read[83] et al. in China. Both surveys show a power-law relationship between distance from home and number of contact events.

The questionnaire does not include contact age, eliminating the possibility of exploring potential differences in age-structured mixing patterns in different social contexts.

A novel approach to sampling the population for social contact surveys has been implemented by Stein et al in both Thailand and the Netherlands [76, 85]. The process is used to assess socio-demographic correlation between contacts. The approach initializes a 'seed' sample. Following completion of the questionnaire, seed participants are requested to recruit a small sub-set of their own contacts, who then repeat the process until no more participants are recruited or a maximum number of cycles are completed. This theoretically has potential for a deeper understanding of types of individuals who mix together. Analysis suggests that in general recruits were assortative by age, education level, household size and vaccination preference [76]. However, there are indications of high recruitment rates from within participants' households.

This approach is limited in two ways. Firstly, only a small proportion of participants were successful in recruiting their contacts making the sample size for comparison small. Secondly, accurate representation of contacts relevant to infectious disease transmission

may be affected by the “friendship paradox”: an individual’s contacts have more contacts than them on average [89]. Qualitative similarities between recruits and contacts have been demonstrated, but this does not necessarily support a proposal that these covariates are also predictive of contacts relevant to the transmission of infectious disease.

There are a number of qualitative consistencies between the findings of all diary-based surveys, for example age assortativity and relationship with household composition. There are however significant differences in specific measures of number of contacts and the composition of age structured mixing. This is most clear from comparison of the constituent countries of the Polymod study, which used a consistent sampling and survey methodology in each setting. Analysis has shown that these differences are clear enough to result in variance in outbreak dynamics and the relative success of identical vaccination programs[14].

Although diary-based contact surveys have proven successful in identifying differences in contact rates between age groups, none has yet attempted to quantify differences in contact rates between social sub-groups of the population explicitly. Sociological studies have sought to measure social integration of socio-economic and ethnic groups in the UK[90, 91], but the measures of integration cannot be directly applied to the transmission of infectious disease.

Three clear limitations within contact survey methodology make their use for understanding widespread interaction between social groups challenging within this framework:

1. *Participant burden*: In general, longer surveys have reduced participant response rates. Hence, by increasing complexity of information gathered through a survey both the scale and representation of the sample can suffer [92, 93].
2. *Statistical power*: As the aggregation of the population becomes more specific, for example, introducing ethnic background or social status of participants, the proportion of the population in these groups reduces. This requires either more targeted recruitment methods or increased numbers surveyed. This can greatly increase the cost and effort required.
3. *Participant interpretation*: The definition of social groups, even by ethnicity, can vary significantly, particularly between those of different cultural heritage. This makes it difficult to ask participants to assign a social group to contacts in a consistent way.

Other methods of recording social contacts

There have been multiple studies that make use of proximity sensors to record interaction between members of a population[94–98]. These are powerful techniques for measuring complete contact networks within small, closed populations such as school classes, however they require the entire population to be included to record every contact made between its members, which becomes untenable in larger populations.

Data driven parameterisation

Another way to establish interaction between groups or individuals is to analyse secondary data sets that give some information about population structure or behaviour.

Models with geographically aggregated populations can use various data on interaction between those sites in conjunction with socio-demographic data (ethnic breakdown, socio-economic status, age structure) to aid parameterisation of a model.

One method of parameterising interaction particularly between geographically defined populations is to make assumptions about how populations move relative to location and population size. The models used to estimate spatial interaction are generally known as "spatial kernels". A popular example of a spatial kernel is the Gravity Model [99], which is based on Newtonian gravity. In this kernel mass is replaced with population size. Larger populations interact more strongly than smaller populations and interaction decays with a power-law. This model can be made more flexible by varying exponents on population size and distance between populations. These approximate measures of interaction have been found to perform differently in different settings and at different spatial scales when assessed against movement patterns identified through mobile phone use data, census commuting data and arrival times of pathogen strains[100–102].

Sometimes data is available that gives some explicit information about movement of people. A good example of this is airline flight and passenger data[103], which is freely available for every flight in the world. This has been used fruitfully in assessing risk of transmission between countries[104], the expected arrival time of particular strains of influenza in different countries[105] and assessing the pandemic potential of a novel strain of influenza[106]. This resource is most useful over long distances where flights are the main form of transport, however this assumption breaks down at shorter distances and usually for within country travel, with the exception of larger countries such as the United States of America.

Similarly, government-collected commuting data has been used to parameterise detailed individual-based models [66, 107] of infectious disease transmission. The premise is that individuals generally move between their home and place of work introducing risk of spreading infection between these locations. An important limitation of this data is that it is often presented at a zipcode/postcode area level or similar, which is relatively coarse when considering individual risk of contact, and an assumption must be made about contact within the same work and residential district limiting information about the interaction between social groups explicitly.

Another form of data-driven network, popular for modelling infections in farm animals, is a hub-based network. In these networks data about interaction between 'hubs' (e.g. farms) is explicitly parameterized from, for example, animal freight data[62, 108, 109]. Transmission within the hub can then be assumed to be homogenous and often hubs are used as the agent in the model such that they themselves have a disease status (e.g. Susceptible, Infected, Recovered). This has been shown to be effective in veterinary epidemiology, however this approach has only rarely been used for modelling infection in human populations. One example is for understanding transmission of hospital-acquired infections on a network of health care institutions, parameterised from patient transfer data [110–112].

Models with social structure for understanding inequalities

Despite a lack of explicitly measured differences in contact rates within and between social groups, there have been instances where mathematical models have been used as a means to investigate the inequalities in infectious disease. These models use spatially

aggregated demographic and commuting data to inform rates between socio-economic groups.

Typical approaches include spatially dependent transmission parameters and transport and commuting data [113, 114]. This approach neglects the potential for significant cultural differences in contact behaviour. For example, religious practice, engagement with family members living outside of the household and the nature of work undertaken. These may lead individuals to mix preferentially within their own social sub-group to a greater degree than spatial models predict [115].

The nature of integration of sub-groups is likely to be highly location-dependent even within a country. The relationship between ethnicity, socio-economic status and cultural/religious preference is likely to be complex[91]. It may, however, be possible to identify predictors of integration based on a combination of ecological factors.

Kumar et al modelled Influenza A H1N1 transmission in New Haven County, Connecticut, USA[114]. The analysis used an "Individual based model", a highly detailed simulation designed to qualitatively reflect realistic movements and contact behaviours based on information about the context being modelled. Kumar et al found that by explicitly modelling differences in population structure (age distribution, household size etc.) and commuting patterns between geographical areas, approximately a third of the observed area-level disparities in the 2009 outbreak were reproduced in the model.

Hyder et al performed a similar analysis in Montreal[113], Canada. Interestingly the disparities in Montreal are reversed; individuals of low socio-economic status have lower

risk of infection from influenza. This had previously been demonstrated in a number of ecological studies[116, 117]. The analysis compares models with increasing level of detail regarding the spatial distribution of social deprivation in the City. Results show that spatial distribution of deprivation is an important factor in reproducing the measured disparities. However, the full extent of these disparities could not be explained purely by spatial considerations.

Both of these studies conclude that spatial and demographic elements of population structure are important, but are insufficient to explain the full extent of observed disparities in influenza risk. Other factors that influence mixing patterns beyond population structure and distribution, such as cultural differences and religious practice, may also be important in explaining disparities in transmission. Development of models that include more realistic estimates of mixing within and between social groups is required to better understand how differences in transmission may account for the observed disparities.

Opportunities with school data

As I have discussed, no existing contact surveys specifically capture interaction between social groups and no well-established data-driven approach to parameterisation is well-suited to approximate these interactions. There is a need to identify methodologies to account for social structure effectively in mathematical models. One source of data that, has not yet been utilized in parameterisation of models of infectious disease transmission, is government school data. Both the UK and the Netherlands maintain detailed databases on school children. This data provides the means to study the structure of the school system and hence social structure relevant to infectious disease transmission within the

school-aged population, which is generally accepted to be of high importance in infectious disease dynamics[118]. In this thesis I take advantage of this data to develop frameworks to evaluate transmission within and between particular social groups.

1.4 Aims and objectives

The aim of this thesis is to better understand observed inequalities in infectious disease and uptake of vaccinations. Specifically, to:

A. Improve understanding of how differences in transmission and uptake of vaccination between social groups contribute to observed and previously unexplained transmission dynamics in diverse populations.
B. To advance the description of population structure in relation to contact-patterns and the embedding of these in a modelling framework to explain observed inequalities and to improve understanding of the impact vaccination on inequalities.

To address these aims I have five objectives:

1. Assess the relative importance of contact rate, susceptibility and vaccine uptake on inequalities in infectious diseases of different epidemiological character.

2. Evaluate whether previously observed inequalities during the early phase of the Influenza H1N1 UK outbreak in 2009 are likely to be related to differences in transmission.

3. Develop a framework to quantify social structure within contact networks of school children in a way that can be used in an infectious disease transmission model.

4. Evaluate the potential role of schools in creating inequalities in influenza outbreaks in London.

5. Analyse the impact of faith schools on clustering of children who are susceptible to measles and resultant measles epidemiology in the Netherlands

1.5 Thesis structure

This thesis is presented in nine chapters. Seven of these chapters present the various analyses in 'paper-style', which form the work of the PhD. These are preceded by this introduction and followed by a general discussion of the findings. Each of chapters presents a self-contained piece of work aimed at answering a specific set of questions which pertain to one of the objectives the PhD. The thesis contains four detailed analyses summarised as follows:

Analysis A is found in chapter 2. This analysis aligns with objective 1 and uses a traditional differential equation model to assess impact of social groups with different transmission related behaviour and/or vaccination uptake on inequalities in disease. This is a published research article with multiple authors and is therefore written in first person plural tense. Details of my contribution to the work are found in the cover sheet at the beginning of the chapter.

Analysis B in chapter 3 evaluates the observed inequalities in Influenza H1N1 during the UK outbreak in 2009, which is in line with objective 2.

Analysis C, detailed in chapter 6 uses a novel framework to simulate outbreaks of influenza in school aged children in London in alignment with objective 4.

Analysis D, which spans chapters 7 and 8, this concerns potential clustering of children who are unvaccinated against measles in the Netherlands and the impact of this on measles epidemiology, to address objective 5. Chapter 7 contains network analysis to evaluate clustering of unvaccinated children. Chapter 8 contains simulation studies to investigate the impact of this clustering on measles epidemiology.

In addition to these major analyses, chapters 4 and 5 contain briefer analyses and a detailed description of the two frameworks I developed using national schools' data to investigate social group structure of school aged children, one of which is used in Analyses C and D. These chapters align with objective 3.

1.6 References

1. HM Government UK. **Equality act**. 2010. http://www.legislation.gov.uk/ukpga/2010/15/contents]. Accessed 25 Sep 2019.

2. Heesterbeek H, Anderson RM, Andreasen V, Bansal S, De Angelis D, Dye C, et al. **Modeling infectious disease dynamics in the complex landscape of global health**. *Science (80-)*. 2015, 347:aaa4339–aaa4339. doi:10.1126/science.aaa4339.

3. Dushoff J, Levin S. **The effects of population heterogeneity on disease invasion**. *Math Biosci*. 1995, 128:25–40. doi:10.1016/0025-5564(94)00065-8.

4. Becker NG, Utev S. **The effect of community structure on the immunity coverage required to prevent epidemics**. *Math Biosci*. 1998, 147:23–39.

5. Metcalf CJE, Lessler J, Klepac P, Cutts F, Grenfell DBT. **Impact of birth rate, seasonality and transmission rate on minimum levels of coverage needed for rubella vaccination**. *Epidemiol Infect*. 2012, 140:2290–301. doi:10.1017/S0950268812000131.

6. Geard N, Glass K, Mccaw JM, Mcbryde ES, Korb KB, Keeling MJ, et al. **The effects of demographic change on disease transmission and vaccine impact in a household structured population**. *Epidemics*. 2015, 13:56–64. doi:10.1016/j.epidem.2015.08.002.

7. Hungerford D, Macpherson P, Farmer S, Ghebrehewet S, Seddon D, Vivancos R, et al. **Effect of socioeconomic deprivation on uptake of measles, mumps and rubella vaccination in Liverpool, UK over 16 years: a longitudinal ecological study.** *Epidemiol Infect*. 2016, 144:1201–11. doi:10.1017/S0950268815002599.

8. Hungerford D, Ibarz-Pavon A, Cleary P, French N. **Influenza-associated hospitalisation, vaccine uptake and socioeconomic deprivation in an English city region: an ecological study**. *BMJ Open*. 2018, 8:e023275. doi:10.1136/bmjopen-2018-023275.

9. Panagiotopoulos T, Antoniadou I, Valassi-Adam E, Berger A. **Increase in congenital rubella occurrence after immunisation in Greece: retrospective survey and systematic review How does herd immunity work?**. *BMJ*. 1999, 319:1462–7. doi:10.1136/bmj.319.7223.1462.

10. van den Hof S, Meffre CM, Conyn-van Spaendonck MA, Woonink F, de Melker HE, van Binnendijk RS. **Measles outbreak in a community with very low vaccine coverage, the Netherlands.** *Emerg Infect Dis*. 2001, 7 3 Suppl:593–7. doi:10.3201/eid0707.010743.

11. Cohen BJ, McCann R, van den Bosch C, White J. **Outbreak of measles in an Orthodox Jewish community**. *Wkly releases*. 2000, 4:1675. doi:10.2807/esw.04.03.01675-en.

12. Lernout T, Kissling E, Hutse V, De Schrijver K, Top G. **An outbreak of measles in orthodox Jewish communities in Antwerp, Belgium, 2007-2008: different reasons for accumulation of susceptibles**. *Eurosurveillance*. 2009, 14:19087. doi:10.2807/ese.14.02.19087-en.

13. Fournet N, Mollema L, Ruijs WL, Harmsen IA, Keck F, Durand JY, et al. **Under-vaccinated groups in Europe and their beliefs, attitudes and reasons for non-vaccination; two systematic reviews**. *BMC Public Health*. 2018, 18:196. doi:10.1186/s12889-018-5103-8.

14. Funk S, Knapp JK, Lebo E, Reef SE, Dabbagh AJ, Kretsinger K, et al. **Combining serological and contact data to derive target immunity levels for achieving and maintaining measles elimination**. doi:10.1101/201574.

15. Phelan JC, Link BG, Tehranifar P. **Social Conditions as Fundamental Causes of Health Inequalities: Theory, Evidence, and Policy Implications**. *J Health Soc Behav*. 2010, 51 1_suppl:S28–40. doi:10.1177/0022146510383498.

16. Semenza JC, Giesecke J. **Intervening to Reduce Inequalities in Infections in Europe**. *Am J Public Health*. 2008, 98:787–92. doi:10.2105/AJPH.2007.120329.

17. Semenza JC. **Strategies to intervene on social determinants of infectious diseases**. *Eurosurveillance*. 2010, 15:32–9. doi:10.2807/ese.15.27.19611-en.

18. CSDH. **Closing the gap in a generation**. 2008. doi:10.1080/17441692.2010.514617.

19. Semenza JC, Suk JE, Tsolova S. **Social determinants of infectious diseases: a public health priority.** *Euro Surveill*. 2010, 15:2–4. doi:10.2807/ese.15.27.19608-en.

20. O'Riordan M, Fitzpatrick F. **The impact of economic recession on infection prevention and control**. *J Hosp Infect*. 2015, 89:340–5. doi:10.1016/j.jhin.2014.11.020.

21. Hawker J, Olowokure B, Sufi F, Weinberg J, Gill N, Wilson RC. **Social deprivation and hospital admission for respiratory infection:**. *Respir Med*. 2003, 97:1219–24. doi:10.1016/S0954-6111(03)00252-X.

22. Myles PR, McKeever TM, Pogson Z, Smith CJP, Hubbard RB. **The incidence of pneumonia using data from a computerized general practice database.** *Epidemiol Infect*. 2009, 137:709–16. doi:10.1017/S0950268808001428.

23. Pockett RD, Adlard N, Carroll S, Rajoriya F. **Paediatric hospital admissions for rotavirus gastroenteritis and infectious gastroenteritis of all causes in England: an analysis of correlation with deprivation.** *Curr Med Res Opin*. 2011, 27:777–84. doi:10.1185/03007995.2011.555757.

24. Balasegaram S, Ogilvie F, Glasswell A, Anderson C, Cleary V, Turbitt D, et al. **Patterns of early transmission of pandemic influenza in London - link with deprivation.** *Influenza Other Respi Viruses*. 2012, 6:e35-41. doi:10.1111/j.1750-2659.2011.00327.x.

25. Inglis NJ, Bagnall H, Janmohamed K, Suleman S, Awofisayo A, De Souza V, et al. **Measuring the effect of influenza A(H1N1)pdm09: the epidemiological experience in the West Midlands, England during the "containment" phase.** *Epidemiol Infect*. 2014, 142:428–37. doi:10.1017/S0950268813001234.

26. Zhao H, Harris RJ, Ellis J, Pebody RG. **Ethnicity, deprivation and mortality due to 2009 pandemic influenza A(H1N1) in England during the 2009/2010 pandemic and the first post-pandemic season.** *Epidemiol Infect*. 2015, 143:3375–83. doi:10.1017/S0950268815000576.

27. Sachedina N, Donaldson LJ. **Paediatric mortality related to pandemic influenza A H1N1 infection in England: an observational population-based study.** *Lancet (London, England)*. 2010, 376:1846–52. doi:10.1016/S0140-6736(10)61195-6.

28. Rutter PD, Mytton OT, Mak M, Donaldson LJ. **Socio-economic disparities in mortality due to pandemic influenza in England.** *Int J Public Health*. 2012, 57:745–50. doi:10.1007/s00038-012-0337-1.

29. Tricco AC, Lillie E, Soobiah C, Perrier L, Straus SE. **Impact of H1N1 on socially disadvantaged populations: Summary of a systematic review**. *Influenza Other Respi Viruses*. 2013, 7 SUPPL.2:54–8. doi:10.1111/irv.12082.

30. Nguyen-Van-Tam JS, Openshaw PJM, Hashim A, Gadd EM, Lim WS, Semple MG, et al. **Risk factors for hospitalisation and poor outcome with pandemic A/H1N1 influenza: United Kingdom first wave (May-September 2009).** *Thorax*. 2010, 65:645–51. doi:10.1136/thx.2010.135210.

31. Haroon SMM, Barbosa GP, Saunders PJ. **The determinants of health-seeking behaviour during the A/H1N1 influenza pandemic: an ecological study**. *J Public Health (Bangkok)*. 2011, 33:503–10. doi:10.1093/pubmed/fdr029.

32. Nyland GA, McKenzie BC, Myles PR, Semple MG, Lim WS, Openshaw PJM, et al. **Effect of ethnicity on care pathway and outcomes in patients hospitalized with influenza A(H1N1)pdm09 in the UK.**

Epidemiol Infect. 2015, 143:1129–38. doi:10.1017/S0950268814001873.

33. Greenwood B. **The contribution of vaccination to global health: past, present and future**. *Philos Trans R Soc B Biol Sci*. 2014, 369:20130433. doi:10.1098/rstb.2013.0433.

34. Fine P, Eames K, Heymann DL. **"Herd Immunity": A Rough Guide**. *Clin Infect Dis*. 2011, 52:911–6. doi:10.1093/cid/cir007.

35. Diekmann O, Heesterbeek JAP. **Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation**. *Wiley Ser*. 2000, :322. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471492418.html.

36. Andre F, Booy R, Bock H, Clemens J, Datta S, John T, et al. **Vaccination greatly reduces disease, disability, death and inequity worldwide**. *Bull World Health Organ*. 2008, 86:140–6. doi:10.2471/BLT.07.040089.

37. Masters NB, Eisenberg MC, Delamater PL, Kay M, Boulton ML, Zelner J. **Fine-scale spatial clustering of measles nonvaccination that increases outbreak potential is obscured by aggregated reporting data**. *Proc Natl Acad Sci*. 2020, :202011529. doi:10.1073/pnas.2011529117.

38. Vynnycky E, Gay NJ, Cutts FT. **The predicted impact of private sector MMR vaccination on the burden of Congenital Rubella Syndrome**. *Vaccine*. 2003, 21:2708–19. doi:10.1016/S0264-410X(03)00229-9.

39. Mangtani P, Breeze E, Kovats S, Ng ESW, Roberts JA, Fletcher A. **Inequalities in influenza vaccine uptake among people aged over 74 years in Britain**. *Prev Med (Baltim)*. 2005, 41:545–53. doi:10.1016/j.ypmed.2005.02.001.

40. Mankertz A, Mihneva Z, Gold H, Baumgarte S, Baillot A, Helble R, et al. **Spread of measles virus D4-Hamburg, Europe, 2008-2011.** *Emerg Infect Dis*. 2011, 17:1396–401. doi:10.3201/eid1708.101994.

41. Feder GS, Vaclavik T, Streetly A. **Traveller Gypsies and childhood immunization: a study in east London.** *Br J Gen Pract*. 1993, 43:281–4. http://www.ncbi.nlm.nih.gov/pubmed/8398244.

42. Cuninghame CJ, Charlton CP, Jenkins SM. **Immunization uptake and parental perceptions in a strictly orthodox Jewish community in north-east London**. *J Public Health (Bangkok)*. 1994, 16:314–7. doi:10.1093/oxfordjournals.pubmed.a042990.

43. McDonald R, Ruppert PS, Souto M, Johns DE, McKay K, Bessette N, et al. **Notes from the field: Measles outbreaks from imported cases in Orthodox Jewish communities - New York and New Jersey, 2018-2019**. *Am J Transplant*. 2019, 19:2131–3. doi:10.1111/ajt.15478.

44. Baugh V, Figueroa J, Bosanquet J, Kemsley P, Addiman S, Turbitt D. **Ongoing measles outbreak in Orthodox Jewish community, London, UK**. *Emerg Infect Dis*. 2013, 19:1707–9. doi:10.3201/eid1910.130258.

45. Stein-Zamir C, Abramson N, Shoob H, Zentner G. **An outbreak of measles in an ultra-orthodox Jewish community in Jerusalem, Israel, 2007 - an in-depth report**. *Eurosurveillance*. 2008, 13:5–6. doi:10.2807/ese.13.08.08045-en.

46. Harmsen IA, Ruiter RAC, Paulussen TGW, Mollema L, Kok G, de Melker HE. **Factors that influence vaccination decision-making by parents who visit an anthroposophical child welfare center: a focus group study.** *Adv Prev Med*. 2012, 2012:175694. doi:10.1155/2012/175694.

47. Ruijs WLM, Hautvast JLA, van IJzendoorn G, van Ansem WJC, van der Velden K, Hulscher ME. **How**

orthodox protestant parents decide on the vaccination of their children: a qualitative study. *BMC Public Health*. 2012, 12:408. doi:10.1186/1471-2458-12-408.

48. Woudenberg T, van Binnendijk RS, Sanders EAM, Wallinga J, de Melker HE, Ruijs WLM, et al. **Large measles epidemic in the Netherlands, May 2013 to March 2014: changing epidemiology.** *Euro Surveill*. 2017, 22. doi:10.2807/1560-7917.ES.2017.22.3.30443.

49. Velzen E van, Coster E de, Binnendijk R van, Hahné S. **Measles outbreak in an anthroposophic community in The Hague, The Netherlands, June-July 2008**. *Eurosurveillance*. 2008, 13:18945. doi:10.2807/ese.13.31.18945-en.

50. Van Lier EA, Oomen PJ, Giesbers H, Conyn-van Spaendonck MAE, Drijfhout IH, Zonnenberg-Hoff IF, et al. **Vaccinatiegraad Rijksvaccinatieprogramma Nederland Verslagjaar 2014**. 2014. www.rivm.nl. Accessed 5 Dec 2019.

51. Goldlust S, Lee E, Bansal S. **Assessing the distribution and drivers of vaccine hesitancy using medical claims data**. In: ISDS 2016 Conference Abstracts. 2016. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5462169/pdf/ojphi-09-e012.pdf. Accessed 16 Sep 2019.

52. Brauer F. **Mathematical epidemiology: Past, present, and future**. *Infect Dis Model*. 2017, 2:113–27. doi:10.1016/j.idm.2017.02.001.

53. Ross R, Hudson HP. **An Application of the Theory of Probabilities to the Study of a priori Pathometry. Part III**. *Proc R Soc A Math Phys Eng Sci*. 1917, 93:225–40. doi:10.1098/rspa.1917.0015.

54. Kermack WO, McKendrick AG. **Contributions to the Mathematical Theory of Epidemics. III. Further Studies of the Problem of Endemicity**. *Proc R Soc A Math Phys Eng Sci*. 1933, 141:94–122. doi:10.1098/rspa.1933.0106.

55. Kermack WO, McKendrick AG. **A Contribution to the Mathematical Theory of Epidemics**. *Proc R Soc A Math Phys Eng Sci*. 1927, 115:700–21. doi:10.1098/rspa.1927.0118.

56. Kermack WO, McKendrick AG. **Contributions to the Mathematical Theory of Epidemics. II. The Problem of Endemicity**. *Proc R Soc A Math Phys Eng Sci*. 1932, 138:55–83. doi:10.1098/rspa.1932.0171.

57. Bernoulli D. **Essai d'une nouvelle analyse de la mortalite causee par la petite verole**. *Mem Math Phys Acad Roy Sci*. 1766, 1.

58. Allen LJS. **An Introduction to Stochastic Epidemic Models**. In: Brauer F, van den Driessche P, Wu J, editors. Mathematical Epidemiology. Lecture Notes in Mathematics, vol 1945. Berlin, Heidelberg: Springer; 2008. p. 81–130. doi:10.1007/978-3-540-78911-6_3.

59. Anderson RM, May RM. **Vaccination and herd immunity to infectious diseases**. *Nature*. 1985, 318:323–9. doi:10.1038/318323a0.

60. Anderson RM, May RM. **Age-related changes in the rate of disease transmission: Implications for the design of vaccination programmes**. *J Hyg (Lond)*. 1985, 94:365–436. doi:10.1017/S002217240006160X.

61. Bolker BM, Grenfell BT. **Impact of vaccination on the spatial correlation and persistence of measles dynamics.** *Proc Natl Acad Sci*. 1996, 93:12648–53. doi:10.1073/pnas.93.22.12648.

62. Keeling MJ, Danon L, Vernon MC, House TA. **Individual identity and movement networks for disease metapopulations**. *Proc Natl Acad Sci*. 2010, 107:8866–70. doi:10.1073/pnas.1000416107.

63. Riley S. **Large-Scale Spatial-Transmission Models of Infectious Disease**. *Science (80-)*. 2007, 316:1298–301. doi:10.1126/science.1134695.

64. Keeling MJ, Eames KTD. **Networks and epidemic models.** *J R Soc Interface*. 2005, 2:295–307. doi:10.1098/rsif.2005.0051.

65. Eames KTD, Read JM, Edmunds WJ. **Epidemic prediction and control in weighted networks.** *Epidemics*. 2009, 1:70–6. doi:10.1016/j.epidem.2008.12.001.

66. Grefenstette JJ, Brown ST, Rosenfeld R, DePasse J, Stone NTB, Cooley PC, et al. **FRED (a Framework for Reconstructing Epidemic Dynamics): an open-source software system for modeling infectious diseases and control strategies using census-based populations.** *BMC Public Health*. 2013, 13:940. doi:10.1186/1471-2458-13-940.

67. Wallinga J, Teunis P, Kretzschmar M. **Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents.** *Am J Epidemiol*. 2006, 164:936–44. doi:10.1093/aje/kwj317.

68. Edmunds WJ, O'Callaghan CJ, Nokes DJ. **Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections.** *Proc Biol Sci*. 1997, 264:949–57. doi:10.1098/rspb.1997.0131.

69. Hoang T, Coletti P, Melegaro A, Wallinga J, Grijalva CG, Edmunds JW, et al. **A Systematic Review of Social Contact Surveys to Inform Transmission Models of Close-contact Infections**. *Epidemiology*. 2019, 30:723–36. doi:10.1097/EDE.0000000000001047.

70. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. **Social contacts and mixing patterns relevant to the spread of infectious diseases.** *PLoS Med*. 2008, 5:e74. doi:10.1371/journal.pmed.0050074.

71. Ibuka Y, Ohkusa Y, Sugawara T, Chapman GB, Yamin D, Atkins KE, et al. **Social contacts, vaccination decisions and influenza in Japan.** *J Epidemiol Community Health*. 2016, 70:162–7. doi:10.1136/jech-2015-205777.

72. Fu Y chih, Wang D-WW, Chuang J-HH. **Representative Contact Diaries for Modeling the Spread of Infectious Diseases in Taiwan**. *PLoS One*. 2012, 7:e45113. doi:10.1371/journal.pone.0045113.

73. Béraud G, Kazmercziak S, Beutels P, Levy-Bruhl D, Lenne X, Mielcarek N, et al. **The French Connection: The First Large Population-Based Contact Survey in France Relevant for the Spread of Infectious Diseases.** *PLoS One*. 2015, 10:e0133203. doi:10.1371/journal.pone.0133203.

74. Hens N, Goeyvaerts N, Aerts M, Shkedy Z, Van Damme P, Beutels P. **Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium.** *BMC Infect Dis*. 2009, 9:5. doi:10.1186/1471-2334-9-5.

75. Danon L, Read JM, House TA, Vernon MC, Keeling MJ. **Social encounter networks: characterizing Great Britain.** *Proc Biol Sci*. 2013, 280:20131037. doi:10.1098/rspb.2013.1037.

76. Stein ML, van der Heijden PGM, Buskens V, van Steenbergen JE, Bengtsson L, Koppeschaar CE, et al. **Tracking social contact networks with online respondent-driven detection: who recruits whom?**. *BMC Infect Dis*. 2015, 15:522. doi:10.1186/s12879-015-1250-z.

77. Dodd PJ, Looker C, Plumb ID, Bond V, Schaap A, Shanaube K, et al. **Age- and Sex-Specific Social Contact Patterns and Incidence of Mycobacterium tuberculosis Infection.** *Am J Epidemiol*. 2016,

183:156–66. doi:10.1093/aje/kwv160.

78. Willem L, Van Kerckhove K, Chao DL, Hens N, Beutels P. **A nice day for an infection? Weather conditions and social contact patterns relevant to influenza transmission.** *PLoS One*. 2012, 7:e48695. doi:10.1371/journal.pone.0048695.

79. Ajelli M, Litvinova M. **Estimating contact patterns relevant to the spread of infectious diseases in Russia**. *J Theor Biol*. 2017, 419:1–7. doi:10.1016/j.jtbi.2017.01.041.

80. Johnstone-Robertson SP, Mark D, Morrow C, Middelkoop K, Chiswell M, Aquino LDH, et al. **Social mixing patterns within a South African township community: implications for respiratory disease transmission and control.** *Am J Epidemiol*. 2011, 174:1246–55. doi:10.1093/aje/kwr251.

81. Kiti MC, Kinyanjui TM, Koech DC, Munywoki PK, Medley GF, Nokes DJ. **Quantifying age-related rates of social contact using diaries in a rural coastal population of Kenya.** *PLoS One*. 2014, 9:e104786. doi:10.1371/journal.pone.0104786.

82. Horby P, Pham QT, Hens N, Nguyen THTHTTY, Le QM, Dang DT, et al. **Social contact patterns in Vietnam and implications for the control of infectious diseases.** *PLoS One*. 2011, 6:e16965. doi:10.1371/journal.pone.0016965.

83. Read JM, Lessler J, Riley S, Wang S, Tan LJ, Kwok KO, et al. **Social mixing patterns in rural and urban areas of southern China.** *Proc Biol Sci*. 2014, 281:20140268. doi:10.1098/rspb.2014.0268.

84. Grijalva CG, Goeyvaerts N, Verastegui H, Edwards KM, Gil AI, Lanata CF, et al. **A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of Peru.** *PLoS One*. 2015, 10:e0118457. doi:10.1371/journal.pone.0118457.

85. Stein ML, van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, van der Heijden PGM, et al. **Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand.** *PLoS One*. 2014, 9:e85256. doi:10.1371/journal.pone.0085256.

86. Leung K, Jit M, Lau EHY, Wu JT. **Social contact patterns relevant to the spread of respiratory infectious diseases in Hong Kong**. *Zimbabwe 21 South Africa*. 1:4–8. doi:10.1038/s41598-017-08241-1.

87. Melegaro A, Del Fava E, Poletti P, Merler S, Nyamukapa C, Williams J, et al. **Social Contact Structures and Time Use Patterns in the Manicaland Province of Zimbabwe**. *PLoS One*. 2017, 12:e0170459. doi:10.1371/journal.pone.0170459.

88. Danon L, House TA, Read JM, Keeling MJ. **Social encounter networks: collective properties and disease transmission**. *J R Soc Interface*. 2012, 9:2826–33. doi:10.1098/rsif.2012.0357.

89. Feld SL. **Why Your Friends Have More Friends Than You Do**. *Am J Sociol*. 1991, 96:1464–77. doi:10.1086/229693.

90. Social Integration Commission. **How integrated is modern Britain?**

91. Finney N, Kapadia D, Peters S. **How are poverty, ethnicity and social networks related?**. 2015. http://www.jrf.org.uk/publications/how-are-poverty-ethnicity-and-social-networks-related.

92. Galea S, Tracy M. **Participation Rates in Epidemiologic Studies**. *Ann Epidemiol*. 2007, 17:643–53. doi:10.1016/j.annepidem.2007.03.013.

93. Bajardi P, Vespignani A, Funk S, Eames KT, Edmunds WJ, Turbelin C, et al. **Determinants of follow-**

up participation in the Internet-based European influenza surveillance platform Influenzanet. *J Med Internet Res*. 2014, 16:e78. doi:10.2196/jmir.3010.

94. Guclu H, Read J, Vukotich CJ, Galloway DD, Gao H, Rainey JJ, et al. **Social Contact Networks and Mixing among Students in K-12 Schools in Pittsburgh, PA**. *PLoS One*. 2016, 11:e0151139. doi:10.1371/journal.pone.0151139.

95. Grantz, H. K, Cummings DAT, Zimmer SM, Vukotich CJ, Galloway DD, Schweizer M Lou, et al. **Age-specific social mixing of school-aged children in a US setting using proximity detecting sensors and contact surveys**. 2020. doi:10.1101/2020.07.12.20151696.

96. Fournet J, Barrat A. **Contact Patterns among High School Students**. *PLoS One*. 2014, 9:e107878. doi:10.1371/journal.pone.0107878.

97. Smieszek T, Castell S, Barrat A, Cattuto C, White PJ, Krause G. **Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants' attitudes.** *BMC Infect Dis*. 2016, 16:341. doi:10.1186/s12879-016-1676-y.

98. Read JM, Edmunds WJ, Riley S, Lesser J, Cummings DAT. **Close encounters of the infectious kind: methods to measure social mixing behaviour**. *Epidemiol Infect*. 2012, 140:2117–30. doi:10.1017/S0950268812000842.

99. Xia Y, Bjørnstad ON, Grenfell BT. **Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics.** *Am Nat*. 2004, 164:267–81. doi:10.1086/422341.

100. Kraemer MUG, Golding N, Bisanzio D, Bhatt S, Pigott DM, Ray SE, et al. **Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings**. *Sci Rep*. doi:10.1038/s41598-019-41192-3.

101. Truscott J, Ferguson NM. **Evaluating the Adequacy of Gravity Models as a Description of Human Mobility for Epidemic Modelling**. *PLoS Comput Biol*. 2012, 8:e1002699. doi:10.1371/journal.pcbi.1002699.

102. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, et al. **On the Use of Human Mobility Proxies for Modeling Epidemics**. *PLoS Comput Biol*. 2014, 10:e1003716. doi:10.1371/journal.pcbi.1003716.

103. Marie M, Meslé I, Hall IM, Christley RM, Leach S, Read JM, et al. **Systematic Review The use and reporting of airline passenger data for infectious disease modelling: a systematic review**. :1. doi:10.2807/1560-7917.ES.2019.24.31.1800216.

104. Flahault A, Deguen S, Valleron A-J. **A mathematical model for the European spread of influenza**. *Eur J Epidemiol*. 1994, 10:471–4. doi:10.1007/BF01719679.

105. Brockmann D, Helbing D. **The Hidden Geometry of Complex, Network-Driven Contagion Phenomena**. *Science (80-)*. 2013, 342:1337–42. doi:10.1126/science.1245200.

106. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. **Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings**. *Science (80-)*. 2009, 324:1557–61. doi:10.1126/science.1176062.

107. Balcan D, Colizza V, Goncalves B, Hu H, Ramasco JJ, Vespignani A. **Multiscale mobility networks and the spatial spreading of infectious diseases**. *Proc Natl Acad Sci*. 2009, 106:21484–9. doi:10.1073/pnas.0906910106.

108. Woolhouse ME., Shaw D., Matthews L, Liu W-C, Mellor D., Thomas M. **Epidemiological implications of the contact network structure for cattle farms and the 20–80 rule**. *Biol Lett*. 2005, 1:350–2. doi:10.1098/rsbl.2005.0331.

109. Bajardi P, Barrat A, Savini L, Colizza V. **Optimizing surveillance for livestock disease spreading through animal movements**. doi:10.1098/rsif.2012.0289.

110. Donker T, Wallinga J, Slack R, Grundmann H. **Hospital Networks and the Dispersal of Hospital-Acquired Pathogens by Patient Transfer**. 2012. doi:10.1371/journal.pone.0035002.

111. Donker T, Wallinga J, Grundmann H. **Patient Referral Patterns and the Spread of Hospital-Acquired Infections through National Health Care Networks**. *PLoS Comput Biol*. 2010, 6. doi:10.1371/journal.pcbi.1000715.

112. Donker T, Henderson KL, Hopkins KL, Dodgson AR, Thomas S, Crook DW, et al. **The relative importance of large problems far away versus small problems closer to home: insights into limiting the spread of antimicrobial resistance in England**. *BMC Med*. 2017, 15:86. doi:10.1186/s12916-017-0844-2.

113. Hyder A, Leung B. **Social deprivation and burden of influenza: Testing hypotheses and gaining insights from a simulation model for the spread of influenza**. *Epidemics*. 2015, 11:71–9. doi:10.1016/j.epidem.2015.03.004.

114. Kumar S, Piper K, Galloway DD, Hadler JL, Grefenstette JJ. **Is population structure sufficient to generate area-level inequalities in influenza rates? An examination using agent-based models**. *BMC Public Health*. 2015, 15:947. doi:10.1186/s12889-015-2284-2.

115. Currarini S, Jackson MO, Pin P. **Identifying the roles of race-based choice and chance in high school friendship network formation.** *Proc Natl Acad Sci U S A*. 2010, 107:4857–61. doi:10.1073/pnas.0911793107.

116. Charland KM, Brownstein JS, Verma A, Brien S, Buckeridge DL. **Socio-Economic Disparities in the Burden of Seasonal Influenza: The Effect of Social and Material Deprivation on Rates of Influenza Infection**. *PLoS One*. 2011, 6:e17207. doi:10.1371/journal.pone.0017207.

117. Crighton EJ, Elliott SJ, Moineddin R, Kanaroglou P, Upshur R. **A spatial analysis of the determinants of pneumonia and influenza hospitalizations in Ontario (1992–2001)**. *Soc Sci Med*. 2007, 64:1636–50. doi:10.1016/j.socscimed.2006.12.001.

118. Worby CJ, Chaves SS, Wallinga J, Lipsitch M, Finelli L, Goldstein E. **On the relative role of different age groups in influenza epidemics.** *Epidemics*. 2015, 13:10–6. doi:10.1016/j.epidem.2015.04.003.

2 Analysis A: Quantifying the impact of social groups and vaccination on inequalities in infectious diseases using a mathematical model

(Published in BMC Medicine: DOI:10.1186/s12916-018-1152-1)

James D. Munday[1,2*], Albert Jan van Hoek[1,2,3†], W. John Edmunds[1,2¶], Katherine E. Atkins[1,2#]

1. Centre for Mathematical Modelling of Infectious Diseases, London School of Hygiene and Tropical Medicine, United Kingdom
2. Department of Infectious Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom
3. National Institute for Public Health and the Environment (RIVM), The Netherlands

Objective: *Assess the relative importance of contact rate, susceptibility and vaccine uptake on inequalities in infectious diseases of different epidemiological character.*

SCHOOL of HYGIENE &TROPICAL MEDICINE

Keppel Street, London WC1E 7HT

T: +44 (0)20 7299 4646
F: +44 (0)20 7299 4656
www.lshtm.ac.uk

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed for each research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH1514285	**Title**	Mr
First Name(s)	James		
Surname/Family Name	Munday		
Thesis Title	The impact of social groups on variation in infectious disease transmission and control		
Primary Supervisor	Dr Albert Jan van Hoek		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

SECTION B – Paper already published

Where was the work published?	BMC Medicine		
When was the work published?	2018		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion			
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	I was responsible for conception of the research question, analysis of data, design of the model, implementation of the model, implementation of the Sobol analysis, interpretation of the results and writing of the publication.

SECTION E

Student Signature	[redacted]
Date	02/12/2019

Supervisor Signature	[redacted]
Date	09-12-2019

Abstract

Background

Social and cultural disparities in infectious disease burden are caused by systematic differences between communities. Some differences have a direct and proportional impact on disease burden, such as health seeking behaviour and severity of infection. Other differences—such as contact rates and susceptibility—affect the risk of transmission, where the impact on disease burden is indirect and remains unclear. Furthermore, the concomitant impact of vaccination on such inequalities is not well understood.

Methods

To quantify the role of differences in transmission on inequalities and the subsequent impact of vaccination, we developed a novel mathematical framework that integrates a mechanistic model of disease transmission with a demographic model of social structure, calibrated to epidemiologic and empirical social contact data.

Results

Our model suggests realistic differences in two key factors contributing the rates of transmission—contact rate and susceptibility—between two social groups can lead to twice the risk of infection in the high risk population group relative to the low risk population group. The more isolated the high risk group, the greater this disease inequality. Vaccination amplified this inequality further: equal vaccine uptake across the two population groups led to up to seven-times the risk of infection in the high risk group. To mitigate these inequalities, the high risk population group would require disproportionately high vaccination uptake.

Conclusion

Our results suggest that differences in contact rate and susceptibility can play an important role in explaining observed inequalities in infectious diseases. Importantly, we demonstrate that, contrary to social policy intentions, promoting an equal vaccine uptake across population groups may magnify inequalities in infectious disease risk.

2.1 Introduction

Reductions in global infectious disease burden have uncovered inequalities in infectious disease health outcomes[1–7]. These inequalities often reflect a disproportionately high incidence observed amongst the most deprived and vulnerable in society[4, 8–10]. Implementing equitable public healthcare relies on prioritizing effective interventions that control the drivers of these inequalities[11].

There may be many contributing factors to inequalities in reported infectious disease health outcomes. Some of these factors have a direct and proportional impact on the relative reported disease burden between social groups; for example, the severity of disease experienced[12, 13], the propensity to seek health care[14] and the reporting rate of disease[15]. In contrast, other factors impact the transmission of infection and these may result in non-linear changes in the relative disease burden between social groups. This latter group of factors include differences in social contact, both within and between social groups, and differences in the susceptibility to infection and infectiousness.

Although indistinguishable when their effects are measured using reported disease burden, these drivers have different implications for delivering equitable public health interventions. For example, in the 2009 H1N1 pandemic disparities in health outcomes between social groups were identified globally. In particular, British Pakistanis had a 3.4 times increased risk of mortality relative to the White British population[16]; many ethnic minority groups (Black, South Asian and South East Asian) had a higher risk (Odds Ratio of 1.33–4.5) of exposure than white Canadians in Ontario[17]; Pacific populations were twice as likely to be exposed to infection than the rest of the New Zealand population[18].

Although these examples would likely present as increased clinical burden in particular sub-groups, the drivers of these differences are difficult to determine. Even though the results from New Zealand indicate differences in transmission rate between sub-groups, the seroprevalence data do not provide enough information to identify the specific driver responsible.

Vaccination is an important intervention in infectious disease control because it reduces disease burden in those vaccinated as well as reducing onward transmission to unvaccinated people. The strength of this indirect protection non-linearly depends on the transmission rate[19]. Therefore, if inequalities are caused by differences in transmission between social groups, vaccination may benefit some groups more than others. The impact of vaccination on inequality in infectious disease outcome is therefore unclear.

To address this gap in our knowledge, we developed a novel mathematical model of the transmission of two vaccine-preventable infections circulating in a population with two social groups characterised by different transmission properties. To quantify the effect of differences in transmission on disease inequality between the social groups, we parameterised the model using realistic estimates of susceptibility and contact structure informed by empirical social mixing data. Using our model, we investigated how the overall impact of vaccination is distributed between two social subgroups, and the effect on inequality in disease incidence.

In addition, we determined the optimal vaccine allocation needed to eliminate inequality.

2.2 Methods

We developed a novel mathematical model to evaluate whether differences in contact rate and the susceptibility to infection between two social groups can explain disease inequality across a population. We used this model to quantify how a vaccination programme affects these inequalities. Our mathematical model combined a dynamic epidemiological model of disease transmission with an age-structured population model of two distinct social groups (Figure 2.1).

Population model

To simulate the demographics of a high-income country, we modelled a stable age distribution with birth rate equal to death rate, a life expectancy of 80 years, and mortality only occurring after 70 years of age at a constant rate. The population model was stratified into n_{age}=15 age groups (0–4, 5–9, … , 65–69, 70+ years) with continuous ageing between age groups. The age-structured population model was further stratified into two social groups of equal size, with the same proportion male and female and identical age structure. Throughout the paper, the social groups with high and low transmission are labelled group *H* and group *L*, respectively.

Epidemiological model

Our dynamic transmission model tracked the proportion of the population as susceptible (*S*), infected but not infectious (*E*), infectious (*I*), and permanently immune to infection (*R*) (Figure 2.1A).

The transmission between and within the two social groups was captured by three mechanisms. The first two control the underlying differences between the two social groups that are potential drivers of inequality: i) a difference in contact intensity between the two groups, expressed as the relative rate at which members of group *L* interact with members of their own group, compared to the rate at which members within group *H* interact with one another ('Contact intensity', $0 < \chi < 1$), and ii) a difference in susceptibility to infection, expressed as the relative susceptibility for members of group *L* compared to members of group *H* ('Susceptibility', $0 < \eta < 1$). The third mechanism determines the integration of the two social groups: iii) the relative rate at which individuals from one social group contact members of the opposite social group ('Integration', $0 < \xi < 1$). For example, ξ = 0.15 corresponds to contact between group *H* and group *L* at 15% of the rate of contact within group *H*. The rate of contact between the groups remained symmetrical i.e. the rate of contact from group *H* to group *L* was the same as the rate of contact from group *L* to group *H*. The force of infection, λ, for the susceptible population in age group i and social groups *H* or *L* is therefore dependent on the social group specific susceptibility, the age- and social group-specific contact rate, and the reproductive number, R_0, of the disease (Figure 2.1C) and can be expressed as:

$$\lambda_{i,H} = \sum_{j=1}^{15} r\beta_{ij}\left(I_{j,H} + \xi I_{j,L}\right) \quad (1a)$$

$$\lambda_{i,L} = \sum_{j=1}^{15} r\eta\beta_{ij}\left(\xi I_{j,H} + \chi I_{j,L}\right) \quad (1b)$$

Where β_{ij} is the age-specific transmission rate from age group j to age group i and $I_{j,H}$ and $I_{j,L}$ are the proportion infectious in age group j and social group *H* and *L* respectively.

To keep R_0 constant when the relative contact rate (χ), susceptibility (η) and integration (ξ) of the social groups were changed, we scaled the force of infection using a linear operator, r. This approach allows parameters of interest (relative contact rate (χ), susceptibility (η) and integration (ξ) and R_0) to be varied independently from each other (Supplementary material). All modelling and analysis was performed using Python[20] .

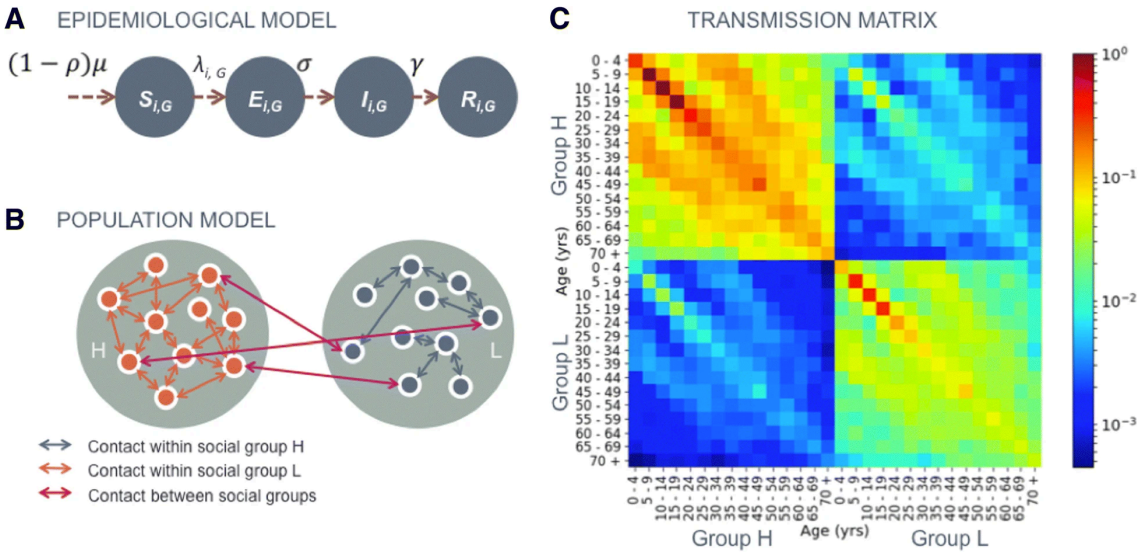


Figure 2.1 Summary of the mathematical model used to quantify inequalities between social groups H (high risk) and L (low risk).

A) The epidemiological model where. $S_{i,G}, E_{i,G}, I_{i,G}. R_{i,G}$ and $\lambda_{i,G}$ are the proportion susceptible, infected not infectious, infectious, recovered and force of infection in age group i and social group G (either group H or group L), ρ is the proportion vaccinated, σ is the rate at which infected individuals become infectious and γ is the rate of recovery from infection, population also moves out of these groups into other age groups and are removed when they die (not shown in this schematic). B) A schematic of the population model with higher contact rate in group H than group L, the groups also differ in susceptibility (not shown). C) An example transmission matrix, showing the relative transmission rate between age and social groups with all social mixing and susceptibility assumptions included with parameterisation $\chi = 0.6, \eta = 0.6, \xi = 0.05$ (rates normalised such that the highest transmission group 10-14 YO in group H has a rate of 1. The same age-group has a rate of 0.36 within group L (Low susceptibility and reduced contact rate, χ and η), 0.05 from group L to group H (between group contact rate, ξ) and 0.03 from group H to group L (between group contact rate and reduced susceptibility, ξ and η).

Parameterisation

Disease scenarios

We parameterised our model for two vaccine-preventable diseases: seasonal influenza and rubella. We quantified the incidence in the total population for both diseases. For influenza, we also quantified the incidence in those aged 60 years and over; who are at risk for severe complications following infection. For rubella, we quantified the incidence in women of childbearing age (WCA) (15-45 years), who serve as a proxy for children born with congenital rubella syndrome after their mothers become infected during pregnancy. The reproduction number, incubation period and infectious period for both diseases were parameterised from literature (Table 2.1). The contact rate between age groups was parameterised with empirical social mixing data collected in the UK arm of the POLYMOD contact survey[21].

Inequality mechanisms

Integration: We informed the parameterisation of ξ, the rate of contact between social groups, relative to the rate of contact within group *H*, using social contact data from the UK arm of the POLYMOD study[21]. We assumed that all household contacts were within their own social group, with a further 70–90% of non-household contacts also within their own social group. The relative rate of contact between social groups, ξ was estimated as 0.05–0.25 (Supplementary material).

Relative contact rate: The feasible range for the contact intensity parameter, χ, the relative rate contact within group *L* compared to group *H*, was also informed by the POLYMOD contact data. For each of the 15 age groups we sorted the participants into quintiles by their household size. We then recombined the age groups, quintile by quintile

to recover five equally sized groups. For each participant, we calculated the total number of contacts from within each person's own social group (using the same assumption as above that all household contacts and 70–90% of non-household contacts were with members of their own social group). The contact intensity parameter, χ was then estimated by evaluating the ratio of the total number of within-group contacts for individuals in every unique pair of quintiles. We estimated the range of ratios as 0.65–0.95 (Supplementary material).

Relative susceptibility: Given the disease specific consideration regarding previous exposure to obtain a parameter for the relative susceptibility (ξ) we investigated the same range of 65–95% susceptibility in group *L* compared to group *H*.

Primary analysis: Quantifying inequalities

The inequality in the population was expressed by the relative risk of infection in the high mixing group (group *H*) relative to the low mixing group (group *L*). We calculated this relative risk across the overall population and for the disease-specific risk groups. For influenza, we calculated the cumulative relative risk over the course of a single outbreak. For rubella, we measured the relative annual infection risk at endemic equilibrium to ensure both rate of transmission and age specific prior exposure to infection were accounted for in our calculation.

	Symbol	Primary analysis	Sobol range**
Population parameters			
Difference in transmission (either*):			
Within group mixing ("Contact")	χ	0.65–0.95	0.65 – 1.54
Relative susceptibility of group *L* to group *H* ("Susceptibility")	η	0.65–0.95	0.65 – 1.54
Quantity of out group mixing relative to within group mixing of group *H* ("Integration")	ξ	0.05–0.25	0.05–0.25
Relative vaccine uptake in group *H* to group *L*	V_H/V_L	1.0	0.70–1.43
* One parameter value set to 1.0 whilst the other adjusted over the 'primary analysis range'' ** Ranges were set so the mid value is the 'base case', which was 1.0 (no difference) for factors which vary group L relative to group H.			
Epidemiological parameters			
Basic reproduction number[22, 23]	R_0		
Influenza		1.8	1.5–4.0
Rubella		6.5	5.0–8.0
Pre-infectious period (days)[24]	σ		
Influenza		2.6	2.6
Rubella		14.0	14.0
Infectious period (days)[24]	γ		
Influenza		4.0	4.0
Rubella		11.0	11.0

Table 2.1 Model parameter values used in base case and sensitivity analyses

Vaccination

For both diseases, we assumed a proportion of individuals become immunised after vaccination—an 'effective coverage'. Consistent with disease-specific immunity profiles, we assumed no waning of vaccine protection over the period of evaluation (lifetime for rubella or one influenza season). Effective coverage for influenza vaccination was identical across all age groups from the beginning of the season; for rubella, vaccine was administered at birth. To allow comparison of results between the two diseases with different R_0 values, we express the effective vaccine coverage as a fraction of the critical vaccination threshold (CVT), $1 - 1/R_0$ i.e. the minimum proportion of the population required to be vaccinated to interrupt transmission. We evaluated the relative risks of infection with no vaccination and with vaccination at 80% of the critical vaccination threshold. Unless otherwise stated, the effective coverage was assumed to be identical between social groups.

Identifying the drivers of inequality

To evaluate the relative importance of the model parameters as drivers for inequality, we used a variance-based global sensitivity analysis, the total Sobol' sensitivity index (S_T)[25, 26], that calculates the proportion of the variance in the relative risk attributable to each parameter and combinations thereof.

2.3 Results

Underlying epidemiology

We ran simulations with no vaccination and no epidemiological differences between group *H* and group *L* (i.e. setting $\chi, \eta, \xi = 1$). We found that the influenza epidemic lasted approximately 21 weeks with a cumulative attack rate of 62% across all age groups and 40% among those over 60 years. For rubella at endemic equilibrium 99.4% of the population were infected before death (95% before the age of 30 years) and the mean age of infection was eight years. The annual incidence among women of childbearing age was 66 per 100,000 (Figure 2.2).

Pre-vaccination inequalities

Influenza

Without vaccination, introducing a relative contact rate of 0.65–0.95 within group *L* compared to group *H*, led to a change in cumulative attack rate in both social groups, and hence a change in the relative risk of infection between the two groups (Figure 2.2). In particular, across base case values of susceptibility and integration, group *H* experienced a relative risk of infection 1.04 - 1.44 compared to group *L* (Figure 2.3A). This relative risk increased to 1.06 - 1.62 (an increase of 1 -12%) among the elderly in group *H*. Less integration between the two groups exacerbated this inequality; when contact between groups was decreased by 67% compared to the base case scenario, ($\xi = 0.05$), the relative risk for group *H* increased to 1.06 - 1.84,(Figure 2.4A).

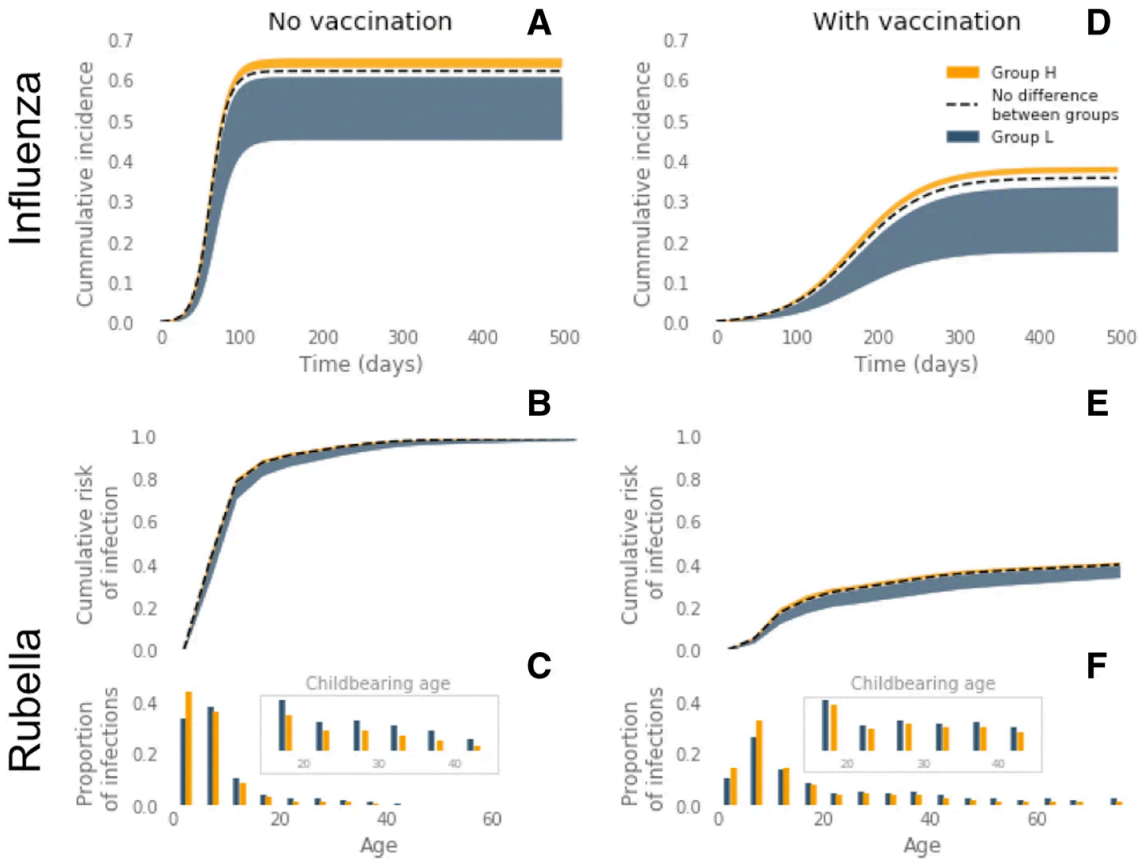


Figure 2.2 Epidemiology predicted by the mathematical model for seasonal influenza and rubella.

With no differences between two population groups (black dashed line) and with differences in susceptibility and contact rate for group H (orange region) and group L (navy region) across feasible range of contact rate within social groups (– and base case values of integration (ξ) and susceptibility (Table 1). A) Cumulative incidence of influenza over a single outbreak with no vaccination. B) Proportion of population infected with rubella by age at endemic equilibrium with no vaccination. C) Proportion of all infections acquired in each 5 year age group, with no vaccination. D) Cumulative incidence of influenza with 37% vaccine uptake (80% of the critical vaccination threshold). E) Proportion of population infected with rubella by age with 67% vaccine uptake (80% of the critical vaccination threshold). F) Proportion of all infections acquired in each 5 year age group, with 67% vaccine uptake (80% of the critical vaccination threshold)

Reducing the susceptibility in group *L* by a factor of 0.65–0.95 relative to group *H*, while maintaining base case values of within-group contact and between group integration, led to 1.05–1.63 times more infections in group *H* than group *L* over the course of the outbreak (Figure 2.3B). Again, the relative risk among the elderly in group *H* was higher

than that of the social group as a whole, with a relative risk of 1.05 - 1.63 under base case assumptions of integration. Relative risk of infection in group *H* increased when the social groups were less integrated relative to the base case scenario to 1.08 - 2.04 ($\xi = 0.05$) and up to 2.49 in the elderly.

Rubella

Unlike our influenza model results, differences in contact rate and susceptibility between the social groups did not result in an inequality in the risk of rubella infection in the whole population (Figure 2.3). However, a more intense contact rate in group *H* or a lower susceptibility in group *L* led to a lower age of infection in group *H* relative to group *L* (Figure 2.2). This difference in the age of infection resulted in a relative risk of infection for women of childbearing age (WCA) in group *H* of 0.64 - 0.95 across feasible ranges of both within-group contact rates and susceptibility. In contrast to the influenza risk group, therefore, our model suggests there is an elevated risk for the low transmission social group (Figure 2.3). Again, in contrast to the influenza model results, varying the level of integration between social groups only marginally affected the relative risk of infection across WCA (Figure 2.4B).

Post Vaccination inequalities

Influenza

Vaccination with a 37% uptake (80% of the critical vaccination threshold) reduced the cumulative attack rate of seasonal influenza from 62% to 30% when transmission in the social groups was identical (Figure 2.2C). However, with differences in contact rate and susceptibility between the two social groups, introducing vaccination increased the inequality between the social groups (Figure 2.3). For example, relative risk of 1.04 - 1.84

before vaccination increased to 1.11 - 2.18 after vaccination with differences in contact rate, and for differences in susceptibility relative risk increased from 1.05 - 2.04 before vaccination to 1.13 - 3.00 after vaccination (with base case integration, $\xi = 0.15$).

Consistent with the results without vaccination, relative risk of infection for group *H* increased when the two social groups were less integrated (Figure 2.4A). When the inequality was driven by feasible changes in either within-group contact rate or susceptibility to infection, the relative risk across the whole of group *H* reached 4.83 and 6.99, respectively, when integration was at its lowest value ($\xi = 0.05$). Therefore, vaccination increased the inequality of disease risk in the social group most at risk of infection by 5 - 241% (Table 2.2).

Although the percentage increase in relative risk after vaccination was less among the elderly in group *H* (5-203%), the relative risk remained higher than in the total population, with a maximum relative risk of 5.19 and 7.52 for differences in contact rate and susceptibility respectively (Figure 2.4A).

The marked increase in inequality in risk of influenza infection as a result of vaccination corresponds to the social group *H* benefiting substantially less from the vaccination programme than group *L*.

Rubella

An effective vaccination uptake of 67% (80% of the critical vaccination threshold) greatly reduced lifetime risk of rubella in both social groups, with less than 40% of the unvaccinated population experiencing infection over their lifetime (Figure 2.2D). With

differences in contact rate between the social groups, vaccination caused an inequality to emerge. Specifically, the relative risk of infection in group *H* relative to group *L* increased from 1.01–1.02 to 1.02–1.42, across a feasible range of within-group mixing patterns (Figure 2.3A). The same result was found as a consequence of susceptibility differences (Figure 2.3B).

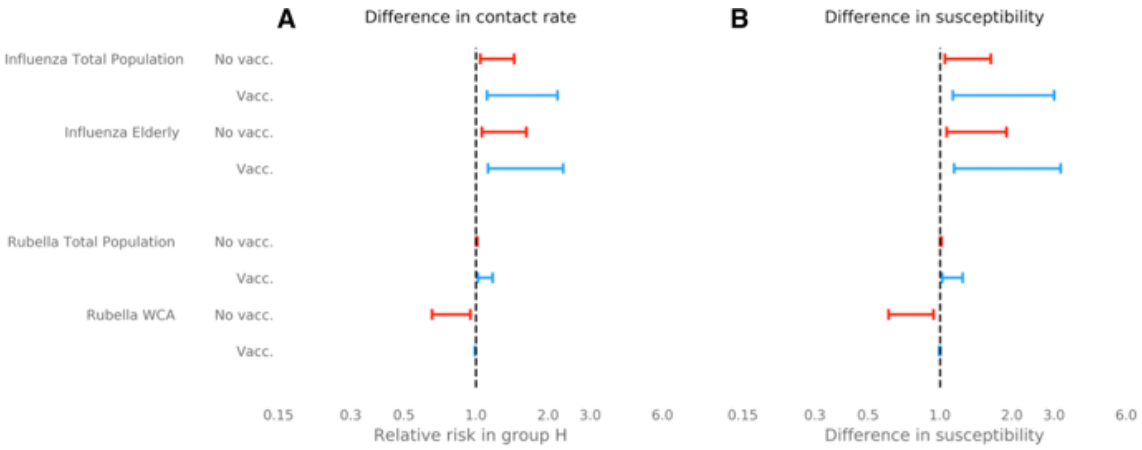


Figure 2.3 Risk of infection in group *H* relative to group *L* in the total population and in risk groups, elderly and women of childbearing age (WCA).

Relative risks shown with no vaccination and vaccination at 80% of critical vaccination threshold (37% for influenza and 67% for rubella). Forest plots show ranges of relative risk for fixed integration of ξ=0.15 and a range of A) ratio of in contact rate in social groups (χ=0.65–0.95) and B) ratio of susceptibility in social groups (η=0.65–0.95)

Furthermore, vaccination reduced the difference in the age of infection between the two social groups (Figure 2.2). The combination of changes in relative risk of infection before death and in age at infection caused a switch in the group most at risk for infection in women of childbearing age. Before vaccination the highest relative risk was among women in group *L* whereas with vaccination the women of childbearing age in the group *H* tend to have a higher risk, with relative risk ranging 0.99–1.16.

Sensitivity analyses show robustness of these results to variation in the relative size and community structure of group L and group H (Figure 16 and Figure 17 of the supplementary material).

Driver of inequality	Infection	Population group	Increase in relative risk
Difference in Contact rate	Influenza	All	4–162%
		Elderly	3–137%
	Rubella	All	2–39%
		WCA	4–72%
Difference in Susceptibility	Influenza	All	5–241%
		Elderly	5–203%
	Rubella	All	2–49%
		WCA	5–86%

Table 2.2 Percentage increase in risk of infection in group *H* relative to group *L* due to vaccination.

Percentage increases in relative risk of infection for the total population (all), women of childbearing age (WCA) and elderly. Results calculated when either the relative within-group contact rates of the two social groups is varied ("Contact" parameter) or when the relative susceptibility of group *L* to group *H* is varied ("Susceptibility" parameter) (Table 1). Integration between the social groups is set at its base case value.

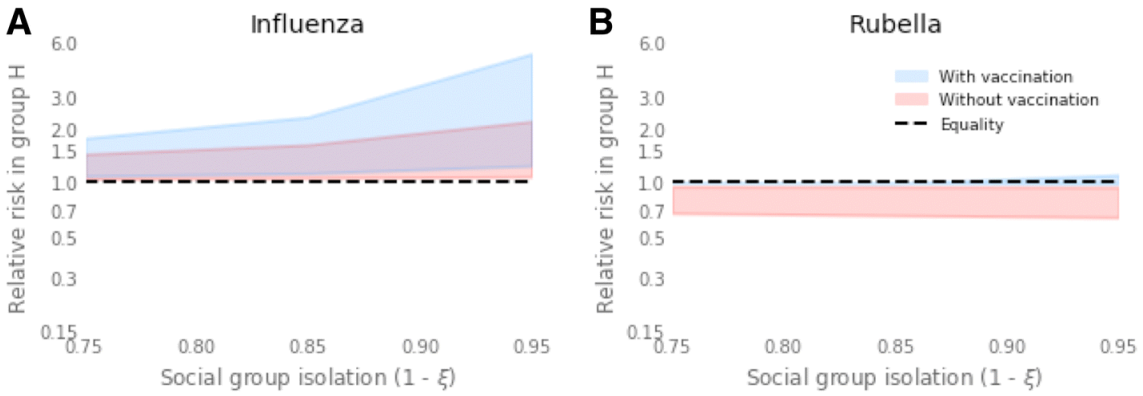


Figure 2.4 Relative risk with social isolation

Full range of relative risk in A) influenza in the elderly (60+ y) and B) rubella in women of childbearing age (15–45 y), due to differences in contact rate ($\chi = 0.6–0.9$) as isolation between sub-groups varies ($\xi = 0.05–0.15$). Red shaded region shows range of relative risk with no vaccination, blue shaded region shows relative risk with vaccination at 80% of the critical vaccination threshold. (37% coverage for influenza, 67% coverage for rubella)

Vaccinating to prevent inequality

By increasing the vaccine uptake in group H relative to group L, the inequalities driven by vaccination, differences in contact rate and differences in susceptibility can be mitigated. To achieve equality in risk of infection for influenza across the entire population, group H had to receive 52–70% of the total number of vaccine doses across the feasible ranges of population parameters (Figure 2.5A). In contrast, small changes in vaccine dose allocation were required to curb inequality in rubella (50.3–52.3%) (Figure 2.5B). The level of integration between the two social groups did not impact the vaccine uptake required in each group to eliminate inequality (results not shown).

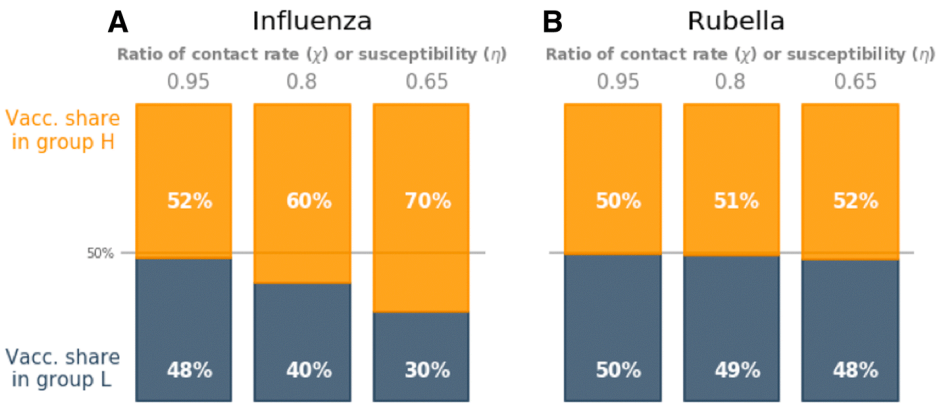


Figure 2.5 Optimal vaccine allocation

Optimal vaccine allocation between social groups required to control disease inequalities in A) influenza and B) Rubella. Results shown for ratio of contact rate in social groups (χ=0.65–0.95) and ratio of susceptibility in social groups (η=0.65–0.95). The total vaccination coverage is 80% of the critical vaccination threshold (37% vaccine uptake for influenza, 67% vaccine uptake for rubella).

Ranking drivers of inequalities

Pre-vaccination era

Without vaccination the magnitude of the inequality (i.e. relative risk of infection for the high transmission group) in influenza was most sensitive to the relative susceptibility of the social groups (S_T = 0.55) and their relative contact rate (S_T = 0.48) (Figure 2.6A). The same was true for rubella (for relative susceptibility S_T = 0.58; for relative within-group contact rate S_T = 0.46). By comparison, sensitivity to integration between the two groups was relatively small, however greater for influenza than rubella (S_T = 0.03 vs. 0.004, respectively) (Figure 2.6B).

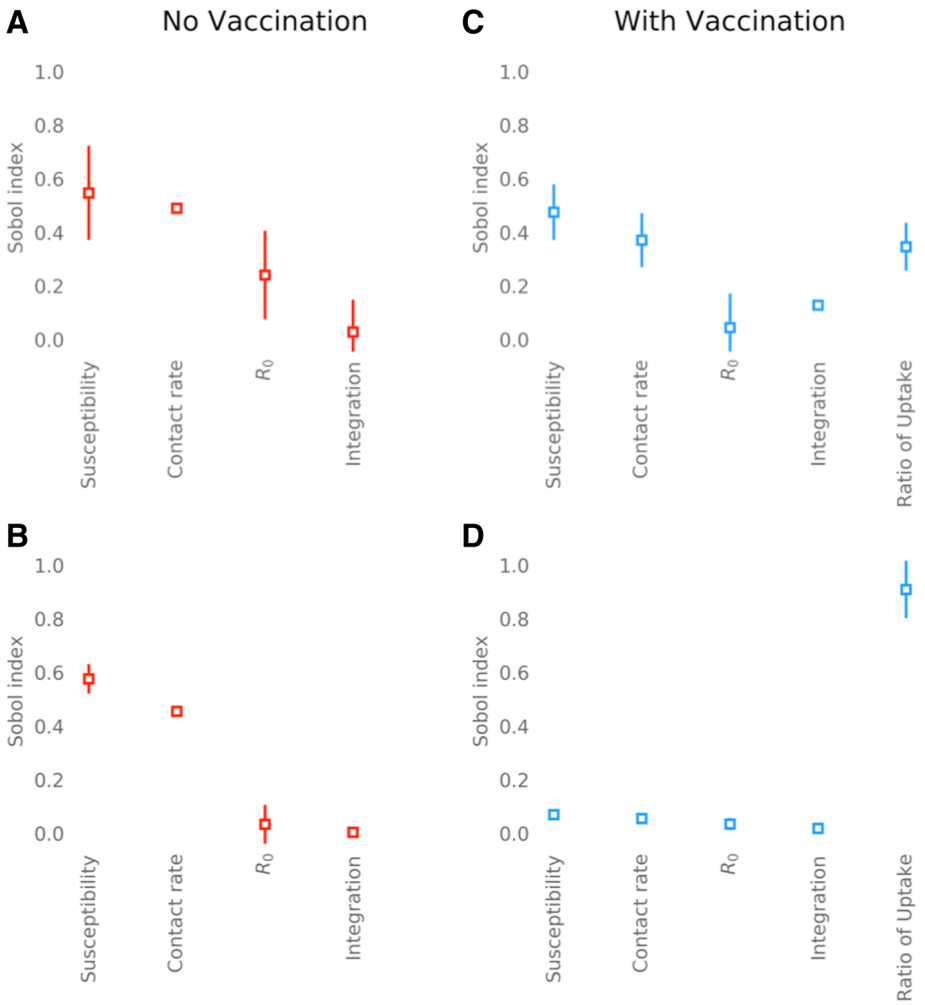


Figure 2.6 Sobol indicies:

Total Sobol' indices, ST , for contact (χ), susceptibility (η), integration (ξ), infectivity (R_0) and difference in vaccination coverage (V_H/V_L) relative risks for rubella and influenza. A) Influenza in the elderly (60+ years) with no vaccination. B) Rubella in women of childbearing age (15–45 years) with no vaccination. C) Influenza in the elderly with vaccination coverage at 37% (80% of the critical vaccination threshold). D) Rubella in women of childbearing age (15–45 years) with vaccination coverage at 67% (80% of the critical vaccination threshold). Error bars show 95% confidence interval.

Vaccination era

Additional variance introduced by differences in vaccine uptake between social groups caused a reduction in the relative sensitivity of inequalities to all other parameters, with the exception of integration. Nonetheless, for influenza, inequality in the disease risk between the two social groups remained most sensitive to relative susceptibility and contact rate (S_T = 0.48; S_T = 0.37). In contrast, inequalities were relatively insensitive to relative vaccine uptake (S_T = 0.35) (Figure 2.6C). Sensitivity to the integration between social groups also increased relative to no vaccination (S_T = 0.13). For rubella relative vaccine uptake between the two social groups had the greatest influence on inequality (S_T = 0.91), diminishing the relative sensitivity of inequality to relative susceptibility and contact rate of the social groups such that they were negligible (Figure 2.6D).

2.4 **Discussion**

Differences in incidence of infectious diseases between social groups have been observed, however the factors that drive these inequalities are not well quantified. Moreover, the impact of vaccination on these inequalities is unclear. We developed a novel mathematical model to simulate influenza and rubella in two connected social groups and assessed the role of differences in two key factors—contact rate and susceptibility—on inequalities as well as the impact of vaccination. Our model suggested that these factors could be responsible for substantial differences in disease epidemiology between social groups. Therefore, these factors may play a significant role in driving observed inequalities in infectious disease outcomes. Furthermore, the results suggest that the impact of these factors on inequalities depend on the characteristics of the pathogen,

as we show that the same differences in transmission are likely to cause greater inequality in influenza than rubella. Vaccination can exacerbate the inequalities even when the uptake is equal between the groups.

These observations have four important implications for public health and immunisation strategies. First, inequality in health is an area of high importance amongst public health authorities[5, 27]. As such there is a appetite for policy that avoids and reduces inequalities in infectious disease outcome[28, 29]. To this end, effort is spent attempting to provide equal distribution of vaccination across social groups in the population[30]. However, our results indicate that equal vaccination uptake could, paradoxically, increase inequalities into high transmission groups, if the vaccine coverage is not high enough to eliminate disease. This result indicates that equal vaccination is not an appropriate measure of equitable intervention, and inequality in disease burden must be evaluated directly.

Second, groups who have social characteristics that place them at a higher risk of infection and who also have a reduced vaccination uptake may be vulnerable to amplified inequalities. For example, during pandemic H1N1 (pH1N1) in 2009, Black and Hispanic populations had a lower uptake of influenza vaccination than the White population in the United States[31]. In addition, countries with self-financed or partially self-paid vaccination programmes may discourage more materially deprived groups from vaccinating; studies[32, 33] in Poland and South Korea have identified lower uptake of vaccination correlates with low Socio-economic status. This leaves the possibility that low uptake may correlate with factors contributing to transmission.

Third, the factors that most influence inequality depend on the underlying disease dynamics and intervention efforts must therefore be disease- and population-specific. For example, our results indicate that differences in vaccine uptake are more important in creating inequalities in rubella than differences in factors associated with transmission rate. This is reflected in the small (0.3–2.3%) change from equal vaccine uptake required to mitigate differences in contact rate or susceptibility (Figure 2.4). However, inequalities in Influenza are more sensitive to differences in transmission related factors than differences in vaccine uptake. This contrast was evidenced when low vaccine uptake in more affluent social groups created "a reversal of health inequalities" with higher prevalence in more affluent areas during a measles outbreak in London, UK in 2001–2002[34]. In contrast, the same geographical region saw a higher attack rate of pH1N1 in more deprived areas[9] only seven years later. This finding suggests that, notwithstanding the potential to increase existing inequalities, for diseases like rubella equal vaccine uptake may be the most practicable target for minimising post vaccination inequalities in disease burden. However, the same approach may not be optimal for Influenza.

Finally, we identified that inequalities resulting from differences in transmission are highly sensitive to the level of integration of subgroups. The importance of integration between social groups becomes more pronounced for diseases with sub-optimal vaccine uptake. This result suggests that inequalities driven by differences in transmission rate or a difference in vaccine uptake may be more likely to occur in highly segregated populations. Our finding could explain inequalities in incidence of infectious disease in urban centres where there is geographical clustering of social and ethnic groups. For example central Birmingham, UK, which was heavily affected by pH1N1 in 2009, is an area where up to 80% of the population is South Asian, an ethnic group associated with

higher risk of transmission[8] . This phenomenon may also contribute to increased risk of outbreaks of measles, often observed in isolated communities with low vaccination coverage[35, 36] . Our findings reinforce the notion that communities that are more isolated should be of particular focus when considering public health strategies for infectious disease. Further, our results highlight the importance of understanding the role of transmission related factors in inequality in populations where social and ethnic groups are becoming more segregated, as inequalities could be set to increase[37].

Our influenza model predicts a relative risk of infection in an unvaccinated population of up to 2.05, within feasible values of social group mixing and susceptibility. This is broadly consistent with data from the pH1N1 epidemic in 2009. For example, a case control study from Ontario, Canada shows that East/Southeast Asian, South Asian and Black Ethnicities had a significantly increased risk of acquiring pH1N1 relative to white Canadians (OR 1.33–4.50)[17]. Similarly in New Zealand a seroprevalence study showed that Pacific Island populations were twice as likely to be infected during the 2009 pandemic than those of European ethnic identity[18]. While there are many examples of observed inequalities in influenza[8, 9, 38–42], studies of inequalities associated with rubella and other endemic childhood infections are often focused on disparities in vaccine uptake rather than disease outcome[43].

While much attention has been given to investigating the impact of transmission heterogeneity on the overall effectiveness of control strategies[44, 45], We build on this work by considering the role of heterogeneity in influencing inequalities in infectious disease outcomes, rather than the overall disease burden. Transmission models have previously been developed to evaluate the impact of social structure on observed

inequalities in reported incidence of pandemic and seasonal influenza[46, 47]. By using socio-economic census data, these studies can replicate some of the location-specific inequalities between pre-defined social groups. However, because it is difficult to disentangle the drivers of inequality underlying these socio-economic groups, the models do not provide a fully generalisable framework in which to evaluate inequality. To overcome this issue, we developed a 'bottom up' approach, in which potential transmission related drivers of inequality are isolated and evaluated. By parameterising our model with empirical social mixing data, we can explicitly capture the contact patterns between age- and social groups and the effect of vaccination. Our generalised framework therefore allows us to disentangle the relative impact of different drivers of inequality and the impact of vaccination on this inequality.

To enable a mechanistic understanding of the drivers of inequality, we made some simplifying assumptions. We assumed that the two social groups in our model have identical age structure and birth rates. It has been shown that differences in age structure and other demographic differences such as birth rate can also result in changes in transmission which lead to inequalities in incidence[46, 47] and the effectiveness of vaccination[48]. To remain consistent with this assumption, we corrected for age distribution when we calculated the range of differences in contact rate between the groups. Furthermore, we assumed gender non-specific contact patterns. In some settings gender differences may exist, particularly in rates of contact between adults and infants[49, 50]. While this gender difference may also differ between social groups, , a recent survey suggest that contact rates between mothers and children are broadly consistent across ethnic and socio-economic groups[50]. Our approach is general and aims to establish the relative impact of various drivers of inequality. As such, our results should not be considered as indicative of the magnitude of specific inequalities, rather the

potential for difference in transmission to explain inequalities and the qualitative nature of the inequalities that may arise from such drivers. We hope the results can be used to target additional analyses at specific scenarios where differences in transmission may arise. For example, where differences in household size distribution or high levels of segregation between social groups prevail.

2.5 Conclusion

Difference in contact behaviour and susceptibility to infection could cause substantial inequality in infectious disease related health outcomes, particularly those related to influenza outbreaks or infections with similar epidemiology. Such inequalities have a highly non-linear relationship with vaccination, which is sensitive to the underlying epidemiology of the infection, ultimately resulting in an increase in inequality after sub-optimal vaccination, even when uptake is equal across the entire population. As such, we advocate measurement of health outcomes rather than vaccination coverage when quantifying the equality of protection across multiple social groups. Moreover, targeted vaccination in known risk groups may reduce overall inequalities in the case of influenza outbreaks, however, due to high sensitivity of rubella inequalities to differences in vaccination coverage, this is not recommended course of action in this case or for similar infections.

2.6 References

1. Millett ERC, Quint JK, Smeeth L, Daniel RM, Thomas SL. **Incidence of community-acquired lower respiratory tract infections and pneumonia among older adults in the United Kingdom: a population-based study.** *PLoS One*. 2013, 8:e75131.

2. Blain AP, Thomas MF, Shirley MDF, Simmister C, Elemraid MA, Gorton R, et al. **Spatial variation in the risk of hospitalization with childhood pneumonia and empyema in the North of England**. *Epidemiol Infect*. 2014, 142:388–98.

3. Myles PR, McKeever TM, Pogson Z, Smith CJP, Hubbard RB. **The incidence of pneumonia using data from a computerized general practice database.** *Epidemiol Infect*. 2009, 137:709–16.

4. Chapman KE, Wilson D, Gorton R. **Invasive pneumococcal disease and socioeconomic deprivation: a population study from the North East of England.** *J Public Health (Oxf)*. 2013, 35:558–69.

5. Semenza JC, Suk JE, Tsolova S. **Social determinants of infectious diseases: a public health priority.** *Euro Surveill Bull Eur sur les Mal Transm = Eur Commun Dis Bull*. 2010, 15:2–4.

6. Semenza JC, Giesecke J. **Intervening to Reduce Inequalities in Infections in Europe**. *Am J Public Heal Semenza Giesecke | Peer Rev | Comment*. 2008, 98.

7. Semenza JC. **Strategies to intervene on social determinants of infectious diseases.** *Euro Surveill Bull Eur sur les Mal Transm = Eur Commun Dis Bull*. 2010, 15:32–9.

8. Inglis NJ, Bagnall H, Janmohamed K, Suleman S, Awofisayo A, De Souza V, et al. **Measuring the effect of influenza A(H1N1)pdm09: the epidemiological experience in the West Midlands, England during the "containment" phase.** *Epidemiol Infect*. 2014, 142:428–37.

9. Balasegaram S, Ogilvie F, Glasswell A, Anderson C, Cleary V, Turbitt D, et al. **Patterns of early transmission of pandemic influenza in London - link with deprivation.** *Influenza Other Respi Viruses*. 2012, 6:e35-41.

10. Jordan R, Verlander N, Olowokure B, Hawker JI. **Age, sex, material deprivation and respiratory mortality**. *Respir Med*. 2006, 100:1282–5.

11. Kawachi I, Subramanian S V, Almeida-Filho N. **A glossary for health inequalities.** *J Epidemiol Community Health*. 2002, 56:647–52.

12. Levy NS, Quyen Nguyen T, Westheimer E, Layton M. **Disparities in the Severity of Influenza Illness: A Descriptive Study of Hospitalized and Nonhospitalized Novel H1N1 Influenza–Positive Patients in New York City: 2009-2010 Influenza Season**. *J Public Heal Manag Pract*. 2013, 19:16–24.

13. Mayoral JM, Alonso J, Garín O, Herrador Z, Astray J, Baricot M, et al. **Social factors related to the clinical severity of influenza cases in Spain during the A (H1N1) 2009 virus pandemic and the CIBERESP Cases and Controls in Pandemic Influenza Working Group, Spain**.

14. Haroon SMM, Barbosa GP, Saunders PJ. **The determinants of health-seeking behaviour during the A/H1N1 influenza pandemic: an ecological study.** *J Public Health (Oxf)*. 2011, 33:503–10.

15. Nyland GA, McKenzie BC, Myles PR, Semple MG, Lim WS, Openshaw PJM, et al. **Effect of ethnicity on care pathway and outcomes in patients hospitalized with influenza A(H1N1)pdm09 in the UK.** *Epidemiol Infect*. 2015, 143:1129–38.

16. Zhao H, Harris RJ, Ellis J, Pebody RG. **Ethnicity, deprivation and mortality due to 2009**

pandemic influenza A(H1N1) in England during the 2009/2010 pandemic and the first post-pandemic season. *Epidemiol Infect*. 2015, 143:3375–83.

17. Navaranjan D, Rosella LC, Kwong JC, Campitelli M, Crowcroft N. **Ethnic disparities in acquiring 2009 pandemic H1N1 influenza: a case–control study**. *BMC Public Health*. 2014.

18. Wilson N, Barnard LT, Summers JA, Shanks GD, Baker MG. **Differential mortality rates by ethnicity in 3 influenza pandemics over a century, New Zealand**. *Emerg Infect Dis*. 2012.

19. Fine P, Eames K, Heymann DL. **"Herd Immunity": A Rough Guide**. *Clin Infect Dis*. 2011, 52:911–6.

20. Python Software Foundation. **Python Language Reference, version 2.7**. Python Software Foundation. 2013.

21. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. **Social contacts and mixing patterns relevant to the spread of infectious diseases.** *PLoS Med*. 2008, 5:e74.

22. Edmunds WJ, Van De Heijden OG, Eerola M, Gay NJ. **Modelling rubella in Europe**. *Epidemiol Infect*. 2000, 125:617–34.

23. Baguelin M, Flasche S, Camacho A, Demiris N, Miller E, Edmunds WJ. **Assessing Optimal Target Populations for Influenza Vaccination Programmes: An Evidence Synthesis and Modelling Study**. *PLoS Med*. 2013.

24. Heymann DL. **Control of communicable disease manual, 20th Edition**. 20th edition. Washington D.C.: American Public Health Association; 2014.

25. Sobol IM. **Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates**. *Math Comput Simul*. 2001, 55:271–80.

26. Herman J, Usher W. **SALib: An open-source Python library for Sensitivity Analysis**. *J Open Source Softw*. 2017.

27. CSDH. **Closing the gap in a generation**. 2008.

28. Hutchins SS, Truman BI, Merlin TL, Redd SC. **Protecting vulnerable populations from pandemic influenza in the United States: A strategic imperative**. American Journal of Public Health. 2009.

29. Blumenshine P, Reingold A, Egerter S, Mockenhaupt R, Braveman P, Marks J. **Pandemic influenza planning in the United States from a health disparities perspective**. Emerging Infectious Diseases. 2008.

30. Peng Y, Xu Y, Zhu M, Yu H, Nie S, Yan W. **Chinese urban-rural disparity in pandemic (H1N1) 2009 vaccination coverage rate and associated determinants: A cross-sectional telephone survey**. *Public Health*. 2013.

31. Uscher-Pines L, Maurer J, Harris KM. **Racial and ethnic disparities in uptake and location of vaccination for 2009-H1N1 and seasonal influenza**. *Am J Public Health*. 2011.

32. Ganczak M, Dmytrzyk-Daniłów G, Karakiewicz B, Korzeń M, Szych Z. **Determinants influencing self-paid vaccination coverage, in 0-5 years old Polish children**. *Vaccine*. 2013.

33. Lee K-C, Han K, Kim JY, Nam GE, Han B-D, Shin K-E, et al. **Socioeconomic Status and Other**

Related Factors of Seasonal Influenza Vaccination in the South Korean Adult Population Based on a Nationwide Cross-Sectional Study. *PLoS One*. 2015.

34. Atkinson P, Cullinan C, Jones J, Fraser G, Maguire H. **Large outbreak of measles in London: reversal of health inequalities.** *Arch Dis Child*. 2005, 90:424–5.

35. Baugh V, Figueroa J, Bosanquet J, Kemsley P, Addiman S, Turbitt D. **Ongoing measles outbreak in Orthodox Jewish community, London, UK.** *Emerg Infect Dis*. 2013, 19:1707–9.

36. Gastañaduy PA, Budd J, Fisher N, Redd SB, Fletcher J, Miller J, et al. **A Measles Outbreak in an Underimmunized Amish Community in Ohio**. *N Engl J Med*. 2016.

37. Casey L. **The Casey Review: A review into opportunity and integration**. London; 2016.

38. Rutter PD, Mytton OT, Mak M, Donaldson LJ. **Socio-economic disparities in mortality due to pandemic influenza in England.** *Int J Public Health*. 2012, 57:745–50.

39. Dee DL, Bensyl DM, Gindler J, Truman BI, Allen BG, D'Mello T, et al. **Racial and Ethnic Disparities in Hospitalizations and Deaths Associated with 2009 Pandemic Influenza A (H1N1) Virus Infections in the United States**. *Ann Epidemiol*. 2011.

40. Quinn SC, Kumar S. **Health inequalities and infectious disease epidemics: a challenge for global health security.** *Biosecur Bioterror*. 2014, 12:263–73.

41. Kumar S, Quinn SC, Kim KH, Daniel LH, Freimuth VS. **The impact of workplace policies and other social factors on self-reported influenza-like illness incidence during the 2009 H1N1 pandemic**. *Am J Public Health*. 2012.

42. Yousey-Hindes KM, Hadler JL. **Neighborhood socioeconomic status and influenza hospitalizations among children: New Haven County, Connecticut, 2003-2010**. *Am J Public Health*. 2011.

43. Doherty E, Walsh B, O'Neill C. **Decomposing socioeconomic inequality in child vaccination: Results from Ireland**. *Vaccine*. 2014.

44. Garnett GP, Anderson RM. **Sexually Transmitted Diseases and Sexual Behavior: Insights from Mathematical Models**.

45. Woolhouse ME, Dye C, Etard JF, Smith T, Charlwood JD, Garnett GP, et al. **Heterogeneities in the transmission of infectious agents: implications for the design of control programs.** *Proc Natl Acad Sci U S A*. 1997, 94:338–42.

46. Kumar S, Piper K, Galloway DD, Hadler JL, Grefenstette JJ. **Is population structure sufficient to generate area-level inequalities in influenza rates? An examination using agent-based models**. *BMC Public Health*. 2015, 15:947.

47. Hyder A, Leung B. **Social deprivation and burden of influenza: Testing hypotheses and gaining insights from a simulation model for the spread of influenza**. *Epidemics*. 2015, 11:71–9.

48. Metcalf CJE, Lessler J, Klepac P, Cutts F, Grenfell DBT. **Impact of birth rate, seasonality and transmission rate on minimum levels of coverage needed for rubella vaccination**. *Epidemiol Infect*. 2012, 140:2290–301.

49. Van Hoek AJ, Andrews N, Campbell H, Amirthalingam G, Edmunds WJ, Miller E. **The Social Life of Infants in the Context of Infectious Disease Transmission; Social Contacts and Mixing Patterns of the Very Young**. 2013.
50. Campbell PT, Mcvernon J, Shrestha N, Nathan PM, Geard N. **Who's holding the baby? A prospective diary study of the contact patterns of mothers with an infant**. *BMC Infect Dis*. 2017, 17.

3 Analysis B: Changing socio-economic and ethnic distribution of cases over the containment phase of the UK Influenza A H1N1 epidemic in 2009 – a comparison of London and Birmingham

Objective: *Evaluate whether previously observed inequalities during the early phase of the Influenza H1N1 UK outbreak in 2009 are likely to be related to differences in transmission.*

3.1 Introduction

During the 2009/10 influenza pandemic, many high-income countries reported a higher incidence of Influenza associated disease in certain social and ethnic sub-groups compared to the rest of the population [1–9]. In the UK, observations of disparities were clearest during the first few months of the outbreak, in the cities of London [10] and Birmingham [11]. Analysis of inequalities in risk of influenza infection can be challenging. Clear measurement of disparity in risk of acquisition of infection requires accurate, detailed data on cases at the point of infection. Without such data, analyses rely on distal measures of infection such as hospitalisation or mortality [2, 4, 7, 8], where reported rates can be influenced by many factors not associated with transmission. Furthermore, aggregating data with low spatial resolution can result in apparent associations between risk and social factors can become exacerbated or diluted due to confounding from geographical variation in risk within regions.

Sufficiently detailed case data is most frequently available from the early stages of an outbreak, chiefly because the small number of cases allows: closer surveillance, higher proportion of cases tested in a laboratory, and more detailed patient records to be kept. However, analysis of data early in an outbreak presents additional challenges. For example, high degree of localisation in early outbreaks and residential clustering of social and ethnic groups geographically may result in measured inequalities simply as a result of the location of index cases. Assessing how disparities in risk change over time might provide additional insight into what is driving them.

During the first three months of the UK Influenza A H1N1 epidemic in 2009, Public Health England (PHE, previously the Health Protection Agency) rolled out reactive antiviral delivery program, in an effort to contain the outbreak by slowing onward transmission through reducing viral load and symptoms in infected individuals [12–15]. As part of this effort, data was collected from all those who reported symptoms and received treatment. Some patients were swabbed and samples sent for laboratory testing. After three months the response and surveillance effort was consolidated into the National Pandemic Flu Service (NPFS), where case data became less detailed [16].

I used the individual level data collected during the initial antiviral delivery program to perform a detailed analysis of the socio-economic and ethnic breakdown of incidence of infection[17]. In particular, I assessed the way in which disparities between socio-economic and ethnic groups develop over the course of the early phase of the outbreak. I compared local outbreaks in the two largest cities in the UK, London and Birmingham, which also accounted for the majority and highest density of cases during the period corresponding to the data I analyse. I used the comparison to identify consistent patterns between the settings, which could provide insight into how disparities my arise, whether there are signs of higher rates of transmission within particular groups, and how social groups may play a particular role in the initiation and determining the dynamics of the early phase of an outbreak.

3.2 Methods

Data overview

The data, collected as part of the initial containment operation accessed from the Fluzone database [17], provides a detailed record of the initial phase of the outbreak across the England and Wales. The individual level data included: Unique identification number, name (anonymised for this analysis), age, date of symptom onset, date that the report arrived at the test centre, full postcode (residence) and school attended. The data also included case status, detailing laboratory testing status (confirmed, test-negative or untested) or whether the reported case is still considered a possible H1N1 infection, or if it has been excluded for another reason (possible, probable or excluded).

To ensure only possible or confirmed cases were analysed, I excluded cases that were coded as either test-negative or 'excluded'. For the purpose of the analysis I required both detailed location and symptom onset time. For cases where no date of symptom onset was recorded but which did include the date received at test centre I estimated the date of symptom onset using the mean time between symptom onset and time received at test centre, as calculated from cases which had both dates recorded. Cases with neither date reported were discarded.

To provide details of socio-economic status and ethnicity breakdown by area, I linked the case data to UK 2011 census data, which I accessed via the Office for National Statistics (ONS) website. Using the full postcode, I assigned each case a Lower Super Output Area (LSOA), which are small geographical regions defined by ONS, with populations of between 800 and 2000 residents. I then linked population data to each case using LSOA level aggregates of the following fields from the 2011 census:

Age distribution: The number of residents of each age (in years) from 0 to 79 and the number of residents 80 and over.

Ethnic group: The number of people who identify as each of the 19 census defined ethnic groups. This data was also broken down by age, which allowed ethnic breakdown to be estimated for children (<=19 yrs) and adults (>19yrs) separately. Ethnic group is coded as shown in Table 3.1.

Deprivation: National Index of Multiple Deprivation (IMD) rank. The rank of the LSOA out of 34,753 LSOAs in England and Wales based on the IMD, a commonly used deprivation measure that captures multiple facets of deprivation including wealth, income, living conditions, quality of life and health outcomes[18].

To assess socio-economic status by relative national and local deprivation, I summarised the deprivation by assigning each LSOA a national decile (the decile (10% band) of the IMD rank in England and Wales). In addition, I identified the local deprivation quintile (the quintile of the IMD rank in the relevant city) of the LSOA in which each case lived.

To classify the ethnic group of each individual case, ethnicity was assigned at an individual level using Onomap software [19], which uses first and last name, prior to anonymisation. The ethnicity classification for this analysis was performed by PHE prior to the commencement of this analysis. Although the inferred ethnicities are based on UK census ethnic groups, the software is not as precise as the census tract presenting a set of broader groups. I have aggregated census groups to the Onomap groups to provide

relevant denominators (Table 3.1). Onomap has been validated in the past [19] and has previously been used for similar analysis of influenza related mortality [20].

ONOMAP Ethnic Group	Census Ethnic Group
White	White: English/Welsh/Scottish/Northern Irish/British
	White: Irish
	White: Gypsy or Irish Traveller
	White: Other White
South Asian	Asian/Asian British: Indian
	Asian/Asian British: Pakistani
	Asian/Asian British: Bangladeshi
Chinese	Asian/Asian British: Chinese
Other Asian	Asian/Asian British: Other Asian
Black	Black/African/Caribbean/Black British: African
	Black/African/Caribbean/Black British: Caribbean
	Black/African/Caribbean/Black British: Other Black
Other/Unclassified	Other ethnic group: Arab
	Other ethnic group: Any other ethnic group
	Mixed/multiple ethnic group: White and Black Caribbean
	Mixed/multiple ethnic group: White and Black African
	Mixed/multiple ethnic group: White and Asian
	Mixed/multiple ethnic group: Other Mixed

Table 3.1 Ethnic Group returned by ONOMAP and the corresponding UK Census codes that were used for population relative population size

Details of the Onomap software and verification

Onomap version 2 (used for the analysis described in this chapter) was developed in 2009 [21]; this version of the software consists of a database of names derived from public registries of over 26 countries and includes 448,657 surnames and 253,881 forenames [19]. The names are each classified by cultural ethnic and linguistic group by evaluating the community structure of the name network (linked by fore- and surnames). The lowest level classification is “Onomap type” which occurs as a community within the name

network, along with this type is a probability score which gives likelihood of a name-type match, which is calculated from the share of the population with the particular name that can be assigned to that type in the training dataset (2001 census). As I stated before, Onomap software assesses both fore- and surname to assign an ethnicity classification. If the assignment conflicts, the software assigns the name-type match with the highest probability score.

Evaluation of the performance of the software has been undertaken by Lakha et. al. [19] in a study which tested the software's performance against multiple datasets in Scotland by comparing assigned ethnicity with parents' country of birth. The Sensitivity, Specificity, positive predictive value and Negative predictive value were reported for each ethnic classification. For clarity I present summary results here in Table 3.2. Notably, the assignment of ethnicity had higher specificity than sensitivity for all ethnicities except British. Complementarily, all non-British ethnicities were assigned with near perfect specificity, whereas British births had relatively poor specificity (68%). I include implications for this variation in sensitivity and specificity in the discussion of this chapter. African ethnicities were assigned with poor sensitivity.

Estimating socio-economic breakdown of cases

To estimate the distribution of cases in London and Birmingham by socio-economic status, I calculated the incidence rate per 100k in each ten-year age group for each national IMD decile. To estimate the distribution of cases by local relative deprivation, I calculated the same for each local IMD quintile.

To summarise the distribution of cases as the outbreak progressed, I plotted Lorentz curves for each week of the outbreak, calculated from the cumulative. The Lorentz plot shows the cumulative proportion of cases by deprivation quintile (i.e. proportion of cases to date in the most deprived 20%, proportion of cases to date in the most deprived 40% etc.). An equal distribution of cases by socio-economic status would return a straight line where the proportion of cases increases by 20% per deprivation quintile; this is called 'the equity line'. Deviation from this line indicates unequal distribution of cases. A Lorentz curve below the equity line indicates a disproportionate share of cases in more affluent areas, whereas a curve above the line indicates disproportionate share of cases in more deprived areas.

Country of birth	*n*	Prevalence (%)	Onomap +ve GROS +ve	GROS +ve	Onomap +ve	Sensitivity% (95% CI)	Specificity% (95% CI)	PPV% (95% CI)	NPV% (95% CI)
British Isles	66,073	86.7	63,795	66,037	67,416	96.6 (96.5–96.7)	64.3 (63.3–65.2)	94.6 (94.5–4.8)	74.4 (73.5–75.3)
Eastern Europe	1616	2.1	1243	1616	1447	76.9 (74.9–7.0)	99.7 (99.7–99.8)	85.9 (84.1–7.7)	99.5 (99.5–97.5)
Accession 8 countries	1413	1.9	1149	1413	1356	81.3 (81.3–81.3)	99.7 (99.7–99.9)	84.7 (82.8–6.7)	99.7 (99.6–99.7)
Poland	1218	1.6	1081	1218	1290	88.8 (87.0–90.5)	99.7 (99.7–99.8)	83.8 (81.8–5.8)	99.8 (99.8–99.9)
South Asia	1636	2.2	1230	1636	2297	75.2 (73.1–77.3)	98.6 (98.5–98.7)	53.6 (51.5–5.6)	99.5 (99.4–99.5)
China	508	0.7	405	508	571	79.7 (76.2–83.2)	99.8 (99.8–99.8)	70.9 (67.2–4.7)	99.9 (99.8–99.9)
Africa	1753	2.3	440	1753	568	25.1 (23.1–27.1)	99.8 (99.8–99.9)	77.46 (74.0–0.9)	98.3 (98.2–98.5)

Table 3.2 Comparison of Onomap results against general register office for Scotland (GROS) birth registration (males and females together). PPV, positive predictive value; NPV, negative predictive value; CI, confidence interval. [Directly from Table 1 in Lakha et. al [19]]

To express unequal distribution of cases by deprivation status in a single value, I calculated the deviation from equity, D, which is the sum of the difference between the Lorentz curve and the equity line at each quintile (or the area between the Lorentz curve and the equity line). Hence a D of 3 would indicate that all the cases were in the most deprived quintile, a D of -3 would indicate that all the cases were in the most affluent quintile. A value of 0 indicates no deviation from equity and cases were distributed equally amongst areas of different deprivation status.

Estimating Ethnic Breakdown of cases using Onomap ethnicity linkage

To estimate the proportion of the population each ethnic group comprised, I aggregated the census ethnic groups as detailed in Table 1.1. To assess the variation in ethnic distribution of cases over the course of the outbreak, I calculated the relative risk at each day of the outbreak based on the cumulative incidence up to that day. As Onomap was not able to attribute an ethnic group to every name in the data, I only included cases that were assigned as one of: White, South Asian, Other Asian, Chinese or Black. Also, in line with this I only calculated the relative proportion of the population amongst only these groups as well. (i.e. proportion of the population that is white is taken as the white proportion of a subset of the total population that is either White, South Asian, Other Asian, Chinese or Black).

I calculated the relative risk of infection in each ethnic group, RR_{et}, as the ratio of the proportion of cases in each group, $P_{cases,et}$, and the proportion of the population each group comprises, $P_{pop,et}$.

$$RR_{et} = \frac{P_{cases,et}}{P_{pop,et}}$$

To ensure the results were not impacted by varying ethnic breakdown by age, I repeated this analysis for the total population and separately for children (<=19 yrs) and adults (>19 yrs).

3.3 Results

Overview of the outbreak

The data comprised 20,301 reported cases nationally, 12,018 of them were remained after exclusions based on case status. There were 1920 with postcodes within Birmingham LSOAs and 3631 in London LSOAs, which also reported either date of symptom onset or date of arrival at the test centre. Of these 855 and 1199 were confirmed in a laboratory for Birmingham and London respectively. The rest remained 'suspected' or 'possible'. Finally, 1,315 and 2,486 had ethnic group successfully inferred using Onomap.

Evaluating the bias of missing data showed no strong correlation between missing postcode information and any variable used in my analysis (Appendix B). There is evidence that in general a higher proportion of samples from white individuals tested negative when submitted for laboratory testing, suggesting that a higher proportion of the un-confirmed cases (possible or suspected) may be false reports (Appendix B).

Age distribution of cases

In Birmingham 72% (70 – 74%, 95% CI) of cases were in children under the age of 19 whereas in London this figure was lower with 60% (58 – 62%, 95% CI) of cases in this

age group. However, there was a higher proportion of cases reported in young adults, between 20 and 29 yrs, 18% (17 – 19%, 95% CI) in London, compared to 11% (10 - 12%, 95% CI) in Birmingham. The majority of cases in both outbreaks were reported in the same 10-week period from the 11th May 2009 (day 130 to day 200) (Figure 3.1). Although there was a higher number of cases in London overall, there was a higher incidence rate in Birmingham.

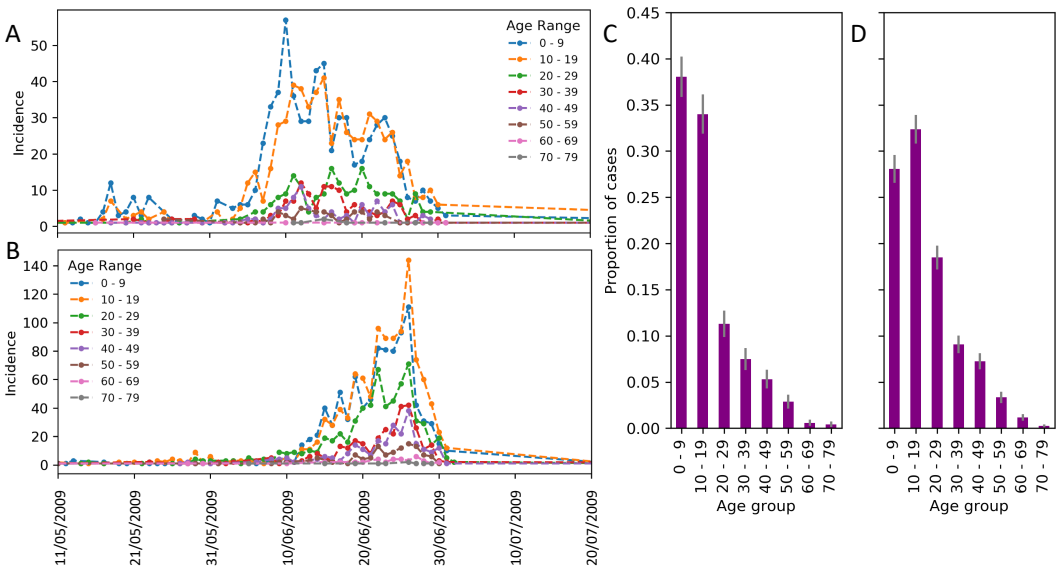


Figure 3.1 Age distribution of cases in Birmingham and London.

Number of cases reported per day between day 130 and 200 in A) Birmingham and B) London stratified by age group. Proportion of cases in each 10 year age group in C) Birmingham and D) London.

Cases by socio-economic status

In Birmingham there was markedly higher incidence in the lower national IMD deciles than the higher deciles, indicating higher incidence in more deprived areas of the city (Figure 3.2). This was replicated for local IMD quintiles where a reduction incidence was clear (Figure 3.3). Incidence per 100k in the most deprived 20% was 2.83 times higher

than the most affluent 20%. This difference is largely due to large disparities in 0 – 19 year-olds where a clear gradient in incidence rate is present and the Rate Ratio between the most deprived and most affluent quintiles was 2.74 (Figure 3.3).

Conversely in London, there was no clear general difference in incidence rate by deprivation, using either national or local grouping of IMD rank. There was, however, slightly higher incidence (1.37 times) in the most deprived 20% than the most affluent 20% overall.

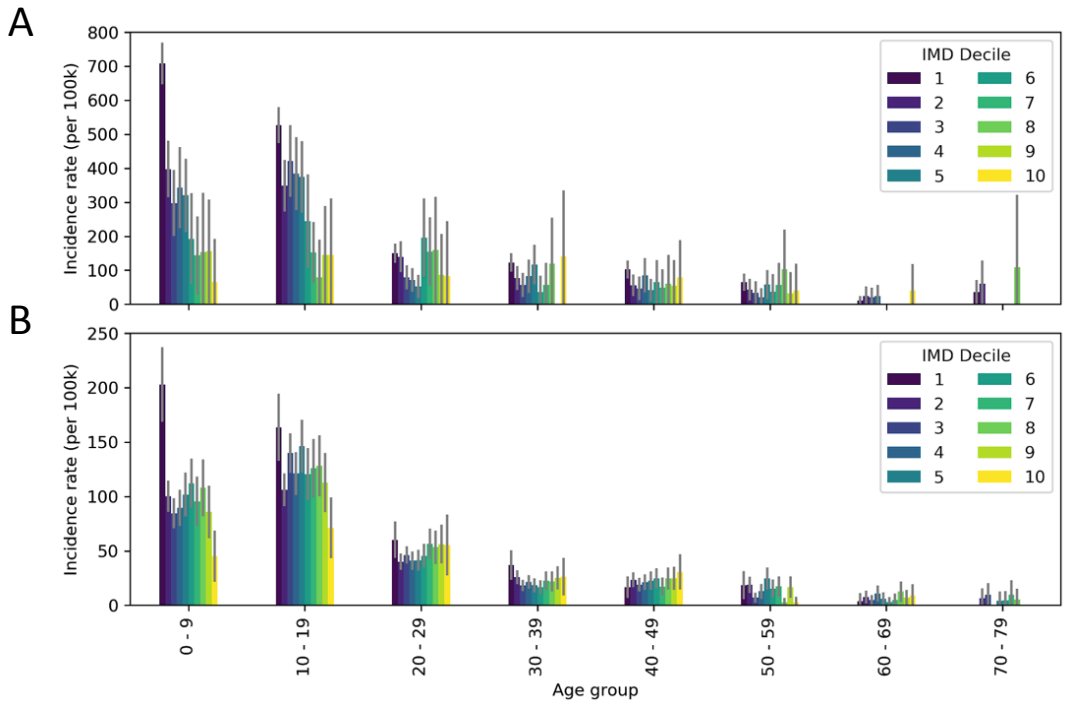


Figure 3.2 Incidence in each 10 year age group per national Index of Multiple Deprivation decile in A) Birmingham and B) London

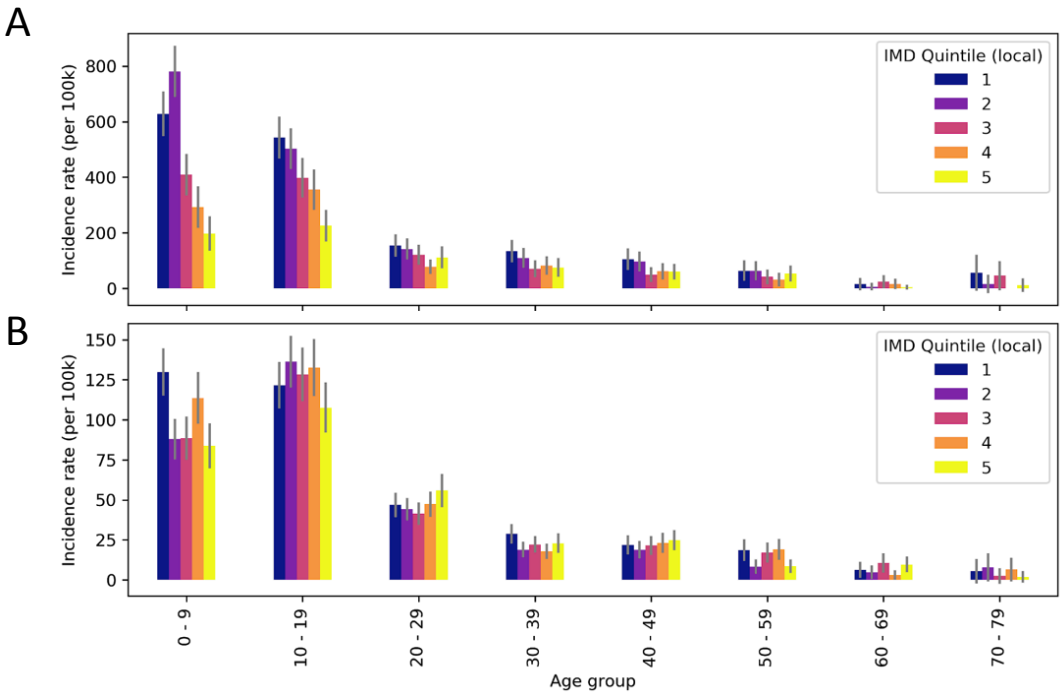


Figure 3.3 Incidence in each 10 year age group per local Index of Multiple Deprivation quintile in A) Birmingham and B) London

The Lorentz curves for Birmingham of cumulative weekly incidence shows that the outbreak largely began in deprived areas. All cases were in areas in the most deprived 40% resulting in a deviation from equity value of 1.6. Cases then spread to more affluent areas gradually, resulting in a deviation from equity of 0.6, and 60% of cases in the most deprived 40% of the population.

In contrast Lorentz curves for London indicate that the outbreak began in more affluent areas with a deviation from equity of -1. The outbreak then progressed to infect more individuals from more deprived areas eventually reaching a deviation from equity of 0.2 (Figure 3.4).

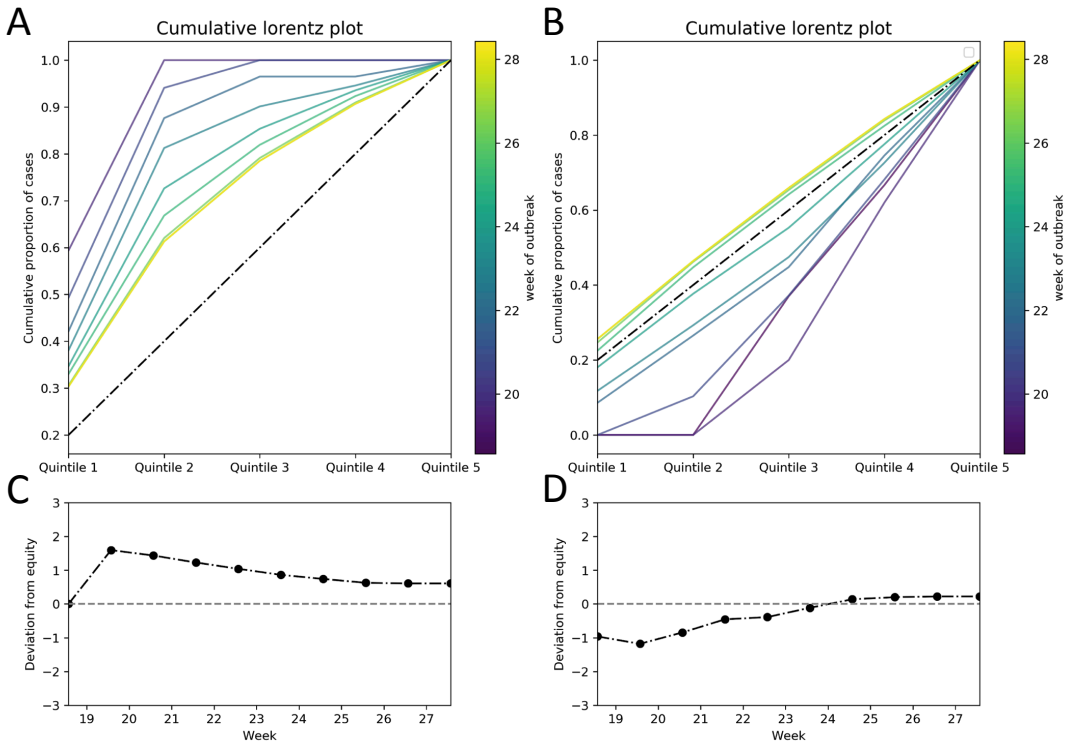


Figure 3.4 Disparities in incidence between local deprivation quintile over time

as: Cumulative incidence by deprivation quintile in each week (Lorentz plot) in A) Birmingham and B) London; Deviation from equity for A) Birmingham and B) London.

Cases by ethnic group

In Birmingham the majority of cases were in individuals identified by Onomap as White (37%) or South Asian (43%). The proportion of the population that is South Asian, however is substantially lower, which results in a relative risk of infection in South Asians of 1.89 (1.71 – 2.08, 95% CI) compared to White of 0.64 (0.58 – 0.72, 95% CI). The relative risk of infection based on the cumulative incidence at each day of the outbreak, reveals that the outbreak began by infecting mostly white individuals however after day 132 the outbreak progressed into the South Asian population, and disproportionately affected this ethnic group from this point. (Figure 3.5)

In London, the majority of cases were in individuals identified as White by Onomap. South Asians made up 16% of all cases, but South Asians also make up a smaller proportion of the population than in Birmingham. The resulting relative risk in this group was 1.36 (1.22 – 1.51, 95% CI). By assessing the relative risk at each day of the outbreak, it appears that the outbreak initiated in a largely white population. The proportion of cases in South Asians was lower than expected for much of the containment period, however number of cases was small and the confidence intervals straddled 1.0 (Minimum RR was 0.24 (0.032 - 1.72, 95% CI)). However, when incidence increased, a higher proportion of cases were in South Asian individuals, reaching a relative risk in this population of 1.36 (1.22 – 1.51, 95% CI), by the end of the data collection period (Figure 3.5). The relative risks followed similar trend in both adult and child age groups (children <= 19yrs < adults). The relative risk was slightly reduced when stratified by age, however the disparity remained clear (Appendix B).

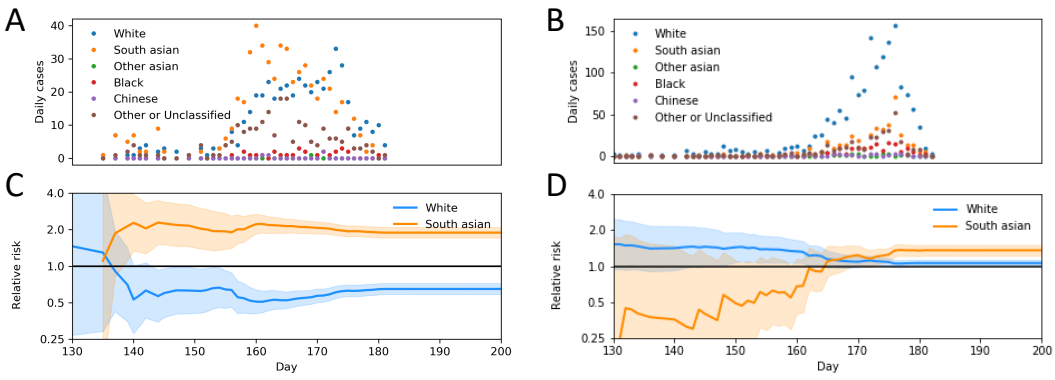


Figure 3.5 Breakdown of reported cases by ethnic group

as determined by ONOMAP. Reported daily incidence of Influenza A H1N1 by ethnic group in A) Birmingham and B) London. Relative risk in White and South Asian ethnic groups in C) Birmingham and D) London; the solid lines show the calculated relative risk (compared to risk of the total population), the shaded areas show the corresponding 95% confidence limits.

3.4 **Discussion**

Disparities in incidence of pandemic influenza between ethnic groups have been reported in multiple high-income settings. Often the analyses that identify such disparities provide a single estimate of relative risk over a defined period of time or using data on particular health outcomes such as hospitalization or mortality. They frequently use ecological analysis to infer relative risk between socio-economic and ethnic groups. By analysing detailed data on cases reported during the initial phase of the UK Influenza A H1N1 epidemic in two urban settings, I have assessed the way disparity in risk develops and varies over time during the early phase of the outbreak.

There were some key differences between the locations, besides those that concern social groups. Higher relative incidence was experienced in Birmingham during the data collection period, however this could have been due to earlier initiation of sustained transmission in the region compared to London. There were also a higher proportion of cases in Children in Birmingham compared to London. However, the similar timing of outbreaks in the two settings and comparable attack rates in the two settings make these well suited to comparison for the purpose of this analysis.

Concerning socio-economic status: There were much clearer differences between deprivation quintiles in Birmingham than in London over the course of the data collection period. These disparities were particularly clear amongst children. However, both outbreaks began with unequal distribution of cases. In London, the majority of cases were in the most affluent region at the beginning of the outbreak. In Birmingham, however,

most cases were in more deprived communities initially. In both cases however, there was gradual movement towards a more equal distribution of cases over the course of the outbreak, suggesting that if an outbreak initiates in one particular socio-economic class it may first progress in that group before dispersing to another. Importantly, the outbreak in London appears to have remained subdued, with low incidence. This suggests that transmission was not sustainable within the population until day 160, after which incidence increases. The increase in incidence coincides with an increase in the proportion of cases reported in more deprived Lower Super Output Areas. In Birmingham, high incidence occurred earlier in the outbreak and mostly in more deprived LSOAs. Incidence then gradually increases in more affluent areas as well.

Disparity in incidence by deprivation status is clearest in children (0 – 19 yrs), who also account for a large proportion of cases in both settings. The proportion of cases in this age group increases greatly as incidence increases overall. This suggests that: outbreaks among children provided the majority of sustained transmission, and differences in transmission within this age group may drive the overall difference in incidence observed by deprivation status.

Concerning ethnic groups, in both the Birmingham and London, an increase in incidence coincided with an increase in the proportion of cases in areas with higher density of South Asians; in Birmingham this occurred around day 132 and in London around day 160. In Birmingham the majority of cases classified were classified by Onomap as South Asian, whereas in London, a smaller, but still disproportionately high proportion of cases were identified as South Asian by Onomap. In both settings the relative risk of infection in South Asians was lower in Children than Adults. It's important to note, however, that the

proportion of cases in South Asians is higher in children than adults, but a higher proportion of children are South Asian overall, resulting in a lower relative risk. One explanation for this is that higher incidence in South Asian children leads to higher incidence in their parents, who represent a smaller proportion of the adult population resulting in higher overall relative risk.

Over the full data collection period, inequality in risk by socio-economic status and ethnic group (specifically higher risk in South Asian individuals) appears to be more substantial in Birmingham than London. However, it should be noted that due to the apparent later initiation of sustained transmission in London, the impact of higher transmission within more deprived regions may not have had time to take its full effect.

The coincidence of increased incidence and presence of cases in the most deprived quintiles, and in South Asian populations could be driven by multiple phenomena. There are two important possible factors, which would be consistent with observations: Firstly, sustained transmission may be more likely to start within more deprived regions and South Asian populations. Secondly, when sustained transmission occurs, transmission rates are generally higher in South Asian and deprived populations. Both of these factors would create inequalities in incidence of infection early on in an outbreak, the first due to higher incidence temporarily in that particular group due to seeding location, the second by transmission rate in a particular group leading to faster accumulation of cases and potentially higher overall incidence. The presence of these dynamics in both setting suggest that if this is the case, a single property that these populations possess that drives this disparity may be present in both cities. Other explanations for this effect include a possibility that the high incidence in South Asians in Birmingham after day 132, initiated

an outbreak in South Asians in London later, through long-range social or familial contact links. This could in turn have led to replication of the observed inequalities at this early stage, when cases were relatively few. There is also the possibility that the observed effects happened purely by chance based on seeding events. However, the relatively large number of cases in London prior to sustained transmission suggests that seeding alone could not explain the observation as there were many opportunities for outbreaks in settings of a different demographic description.

When analysing data collected as part of an initial outbreak response there are always some limitations that give way to potential biases. I would like to highlight two clear limitations here. Firstly, some of the data collected was part of a contact tracing effort[17]. This creates the opportunity for biases to arise, as potential increased surveillance in certain populations may be compounded by active case finding amongst contacts in that population. Both of these may generate a higher case to infection ratio. There is evidence however that a greater proportion of tests resulted in negative result in the White British population, suggesting that case finding and surveillance efforts were not disproportionately focused within the South Asian population. Secondly, there is likely to be substantial variation in testing capacity throughout the initial phase as services become overwhelmed and adapt to demand. This could have the impact of the composition of tests between self-reporting and active case finding varying. There is no clear way to evaluate this in the present case as the source of cases is not recorded. This could impact perceived inequality in risk as actively found cases from known outbreaks could be favoured over self-reported cases and vice versa depending on the priorities of the local response.

The accuracy of ethnicity assignment using Onomap Version 2 must play a role in the interpretation of these results[19]. Since previous verification studies show that Onomap has higher sensitivity when assigning British ethnicity compared to non-British ethnicities and lower specificity assigning British comparted to others (Table 3.2), it is likely that non-British ethnicities was under represented in the assigned data and British ethnicity was over represented. In particular, Black ethnicities may be highly under-represented, with a sensitivity of only 25.1% (23.1%–27.1% 95% CI)[19], compared to other ethnicities which had a sensitivity of over 75%. Other Black ethnicities were not evaluated by Lakha et. al. This poor performance suggests little should be interpreted from the proportion of cases Black ethnicities represent. This is acceptable for this analysis since the major focus is on South Asian and White British ethnicities. When considering this comparison explicitly, the performance in Lakha et. al. suggests that South Asians are more likely to be under represented than over represented in the assigned data, suggesting that estimates for relative risk are conservative. In contrast British ethnicity is likely to be over-estimated, suggesting that disparities may have been even greater than identified in this analysis. Since these potential biases are not expected to vary over time, my analysis of variation in disparities over time should not have been further affected.

The findings of this analysis corroborate previous analyses of inequalities in this phase in the west midlands [11], which used the same dataset. The previous analysis by Inglis et. al. identified a disproportionate incidence in South Asians and more deprived regions of in the West Midlands region overall (of which the city of Birmingham is a part). My analysis builds on these results by evaluating disparities at a finer spatial resolution and analysing how they change over time. Previous analysis of the same period of the

outbreak in London also identify a gradual transition of cases from more affluent groups early in the outbreak, to more deprived communities [10]. However, this analysis used a different data source. My analysis complements this by supporting the findings with a different data set and evaluating disparities by ethnic group in this setting. In addition, by using a national dataset I was also able to compare outbreaks in the two most important settings during the early phase of the outbreak.

To conclude, my detailed analysis of the early phase of the UK Influenza A H1N1 epidemic in 2009 indicates that there may be a connection between the initiation of sustained transmission with the introduction of infection to more deprived areas and South Asian populations, both in Birmingham and London. This phenomenon resulted in higher risk of infection in the most deprived areas and South Asians during the containment phase of the epidemic. This was most clear in Birmingham and particularly in children under the age of 19 years. The combination of higher incidence in children and more pronounced disparities by deprivation status suggest that children may be important in driving inequalities in these settings.

3.5 References

1. Charland KM, Brownstein JS, Verma A, Brien S, Buckeridge DL. **Socio-Economic Disparities in the Burden of Seasonal Influenza: The Effect of Social and Material Deprivation on Rates of Influenza Infection**. *PLoS One*. 2011, 6:e17207. doi:10.1371/journal.pone.0017207.

2. Dee DL, Bensyl DM, Gindler J, Truman BI, Allen BG, D'Mello T, et al. **Racial and Ethnic Disparities in Hospitalizations and Deaths Associated with 2009 Pandemic Influenza A (H1N1) Virus Infections in the United States**. *Ann Epidemiol*. 2011, 21:623–30. doi:10.1016/j.annepidem.2011.03.002.

3. Quinn SC, Kumar S, Freimuth VS, Musa D, Casteneda-Angarita N, Kidwell K. **Racial Disparities in**

Exposure, Susceptibility, and Access to Health Care in the US H1N1 Influenza Pandemic. *Am J Public Health*. 2011, 101:285–93. doi:10.2105/AJPH.2009.188029.

4. Levy NS, Quyen Nguyen T, Westheimer E, Layton M. **Disparities in the Severity of Influenza Illness: A Descriptive Study of Hospitalized and Nonhospitalized Novel H1N1 Influenza–Positive Patients in New York City: 2009-2010 Influenza Season**. *J Public Heal Manag Pract*. 2013, 19:16–24.

5. Navaranjan D, Rosella LC, Kwong JC, Campitelli M, Crowcroft N. **Ethnic disparities in acquiring 2009 pandemic H1N1 influenza: a case–control study**. *BMC Public Health*. 2014, 14:214. doi:10.1186/1471-2458-14-214.

6. Placzek H, Madoff L. **Effect of Race/Ethnicity and Socioeconomic Status on Pandemic H1N1-Related Outcomes in Massachusetts**. *Am J Public Health*. 2014, 104:e31–8. doi:10.2105/AJPH.2013.301626.

7. Mayoral JM, Alonso J, Garín O, Herrador Z, Astray J, Baricot M, et al. **Social factors related to the clinical severity of influenza cases in Spain during the A (H1N1) 2009 virus pandemic**. *BMC Public Health*. 2013, 13:118. doi:10.1186/1471-2458-13-118.

8. Wilson N, Barnard LT, Summers JA, Shanks GD, Baker MG. **Differential mortality rates by ethnicity in 3 influenza pandemics over a century, New Zealand**. *Emerg Infect Dis*. 2012.

9. Yousey-Hindes KM, Hadler JL. **Neighborhood Socioeconomic Status and Influenza Hospitalizations Among Children: New Haven County, Connecticut, 2003–2010**. *Am J Public Health*. 2011, 101:1785–9. doi:10.2105/AJPH.2011.300224.

10. Balasegaram S, Ogilvie F, Glasswell A, Anderson C, Cleary V, Turbitt D, et al. **Patterns of early transmission of pandemic influenza in London - link with deprivation.** *Influenza Other Respi Viruses*. 2012, 6:e35-41. doi:10.1111/j.1750-2659.2011.00327.x.

11. Inglis NJ, Bagnall H, Janmohamed K, Suleman S, Awofisayo A, De Souza V, et al. **Measuring the effect of influenza A(H1N1)pdm09: the epidemiological experience in the West Midlands, England during the "containment" phase.** *Epidemiol Infect*. 2014, 142:428–37. doi:10.1017/S0950268813001234.

12. Longini IM, Halloran ME, Nizam A, Yang Y. **Containing Pandemic Influenza with Antiviral Agents**. *Am J Epidemiol*. 2004, 159:623–33. doi:10.1093/aje/kwh092.

13. Fraser C, Riley S, Anderson RM, Ferguson NM. **Factors that make an infectious disease outbreak controllable**. *Proc Natl Acad Sci*. 2004, 101:6146–51. doi:10.1073/pnas.0307506101.

14. Wu JT, Riley S, Fraser C, Leung GM. **Reducing the Impact of the Next Influenza Pandemic Using Household-Based Public Health Interventions**. *PLoS Med*. 2006, 3:e361. doi:10.1371/journal.pmed.0030361.

15. Black AJ, House T, Keeling MJ, Ross J V. **Epidemiological consequences of household-based antiviral prophylaxis for pandemic influenza**. *J R Soc Interface*. 2013, 10:20121019. doi:10.1098/rsif.2012.1019.

16. Department of Health HGU. **The National Pandemic Flu Service: An Evaluation**. 2011. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/215679/dh_125338.pdf. Accessed 9 Oct 2020.

17. McLean E, Pebody R, Chamberland M, Paterson B, Smyth B, Kearns C, et al. **Epidemiological report**

of pandemic (H1N1) 2009 in the UK. Royal College of General Practitioners.

18. Department for Communities and Local Government, HM Government U. **The English Indices of Deprivation**. 2015. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/465791/English_Indices_of_Deprivation_2015_-_Statistical_Release.pdf. Accessed 25 Feb 2019.

19. Lakha F, Gorman DR, Mateos P. **Name analysis to classify populations by ethnicity in public health: Validation of Onomap in Scotland**. *Public Health*. 2011, 125:688–96. doi:10.1016/J.PUHE.2011.05.003.

20. Zhao H, Harris RJ, Ellis J, Pebody RG. **Ethnicity, deprivation and mortality due to 2009 pandemic influenza A(H1N1) in England during the 2009/2010 pandemic and the first post-pandemic season.** *Epidemiol Infect*. 2015, 143:3375–83. doi:10.1017/S0950268815000576.

21. Mateos P, Longley PA, O'Sullivan D. **Ethnicity and Population Structure in Personal Naming Networks**. *PLoS One*. 2011, 6:e22943. doi:10.1371/journal.pone.0022943.

4 Contact between children – location, duration and frequency of child-to-child contact

Objective: *Develop a framework for to quantify social structure within contact networks of school children in a way that can be used in an infectious disease transmission model.*

4.1 Introduction

Children often contribute disproportionately to transmission of common infections, e.g. measles, rubella and varicella, as well as Influenza. This is demonstrated in Chapter 3 of this thesis and previous analyses [1, 2].

On this basis, the nature of contact between children is of particular interest in the field of infectious disease epidemiology. Understanding contact behaviour of children promises to reap the greatest reward in understanding the overall dynamics of transmission of a large number of important infectious diseases. Happily, the social contact behaviour of children may be more predictable than much of the population. For the most part each child between the age of 4 and 16 spends a large part of the waking time of approximately 70% of their days in an educational setting. A large proportion of the rest of their time is likely to be spent at home. The importance of schools in transmission of infectious disease is evident from the demonstrated impact of holidays and school closure on transmission dynamics during many outbreaks[3–5]. Analysis by Melegaro et. al. [6] suggests that school and home based contacts are most important for transmission of Varicella Zoster Virus and Parvovirus B19. There is also evidence that physical contact, longer duration contacts and more frequent contacts best explain serology of these infections in European countries. The clear importance of schools in transmission has led to a number of targeted social contact surveys and similar studies that aim to understand potentially infectious contact within schools [7–18]. This section details a short analysis where I used existing data from a large social contact survey [19] to assess the relative importance of school and household based contacts in the overall contact network of school-aged children.

4.2 Materials and Methods

Analysis of contact survey data

Survey data

I analysed data collected as part of the Polymod study [19] to establish the proportion of children's contacts that occur either in school or in their household. The contact survey data detail 7290 participants and 97,904 contacts recorded in eight European countries (Belgium, Germany, Finland, Great Britain (GB), Italy, Luxembourg, The Netherlands, and Poland). Participants were asked to record contact events that they experienced over a pre-arranged 24-hour period (between 5 a.m. and 5 a.m. the following day). A contact event was defined either a two-way conversation at least three words in the physical presence of another person or skin-to-skin contact.

Amongst other fields, data contains participant age, contact age, where contact event(s) took place (location(s)), time spent with a single contact (duration), frequency with which the participant and a particular contact typically have meet (frequency) and whether the contact event included physical touch (physical). The possible values of these fields are shown in table 4.1.

Contacts of children

I investigated the locations in which children make contact outside the home by considering the contacts of school-aged children only (4 – 19 yrs.). To assess the role of school contacts I found the proportion of contacts made outside the home in school and in all other places. I also calculated the proportion of contacts made outside the home and not in school that are also associated with a contact event in school or home (i.e. the same

contact recorded in two locations). To account for the time of exposure of contacts I also calculated these proportions for contacts of increasing duration and frequency.

Field	Category
Duration	< 5 mins 5 - 15 mins 15 - 60 mins 1 - 4 hours 4 + hours
Location	Home Work School Leisure Transport Other
Frequency	Daily or almost daily (Daily) About once or twice a week (Weekly) About once or twice a month (Monthly) Less than once a month (A few times a year) For the first time
Contact type	Physical / non-physical

Table 4.1 Categories for the fields in the Polymod contact survey [10] of interest to this analysis

Proportion of total contact time in school and at home

To highlight the relative importance of school and home contacts for school-aged children, I estimated the proportion of contact time in school, at home or other. For individual contacts where contact was made in multiple locations, the survey recorded duration as the sum of contact time over all contact events. Therefore, for such contacts it was necessary for me to infer the proportion of the duration of contact made at home, in school and other. I estimated the ratio of contact durations for contact events in each

location category using contacts with a single location reported (either home, school or other). I used these relative durations to divide the time between locations for contacts with multiple locations reported.

4.3 Results

Age specific contacts at home and elsewhere

In the Polymod social contact survey, 27 % (26.2 - 26.85, 95 % CI) of contacts were reported at home, 73 % (73.2 – 73.8, 95% CI) were reported outside home. In school-aged children the proportion of contacts outside home was higher, 75% (74.9 - 75.8, 95 % CI).

When I filtered the contacts by duration, the proportion of contacts outside the home reduced to 54% (53.8 - 55.0, 95% CI) for contacts with duration of at least 4 hours. The reduction was lower in school-aged children, where the proportion outside the home reduced to 65% (64.1 – 65.8, 95% CI)

Of the total number of outside home contacts 36% (36.2 – 37.0, 95% CI) were reported by children, correcting for population age distribution. For contacts with reported duration of over 4 hours, 58% (57.2 – 58.9, 95% CI) of outside home contacts were reported by children.

Of contacts outside the home 45% (44.1 – 44.9, 95% CI) involved physical touch, whereas 76% (75.9 – 77.0, 95% CI) of home contacts involved touch. Of children's contacts, 51% (50.2 – 51.4, 95% CI) of contacts outside the home involved touch.

Contacts of school-aged children

The majority of contacts of school age children were also school age children (58%, 57.5 – 58.6, 95 % CI). Most of contacts between school age children occur outside of the home 86 % (86.1 – 87.0, 95 % CI). Of contacts out of the home between children, 69% (68.5 – 69.8, 95% CI) of them include an event in school. Of all contacts between school age children, 73% (72.8 – 73.9, 95% CI) of them had contact in school or at home. This proportion increases both when the contacts are filtered for longer duration contacts higher frequency of event (Figure 4.1). Of the events that occurred for the longest duration (>4 hours) with most frequent (at least once a day) contacts, 99.9% (99.9 – 100.0, 95% CI) of contacts between school-aged children and 95% (94.9 – 95.8, 95% CI) of all contact events occurred either in school or at home. Of contact events in locations other than school or home, 23% (22.4 – 23.1, 95% CI) of them also had a contact event recorded in school, home or both. The proportion this increased to over 75% (72.6 – 76.9, 95% CI) for daily contacts with durations of over 4 hours.

Proportion of total contact time at school or at home

89% (88.9 – 89.6, 95% CI) of contacts were recorded in a single location (home, school, other). For contacts where only one location was recorded, the mean duration of a contact event at school, home and other were 222 minutes, 243 minutes and 101 minutes respectively. The average duration for contacts with events recorded both at school and at home was 275 minutes. I weighted duration of contacts' multiple locations recorded according to the overall distribution of contacts in each of school, home, and other locations. I estimated the proportion of contact time reported by children to be 84% (83.0

– 85.4, 95% CI) at school and at home for all contacts and 88% (86.1 – 88.8, 95% CI) at school and at home for contacts reported on weekdays.

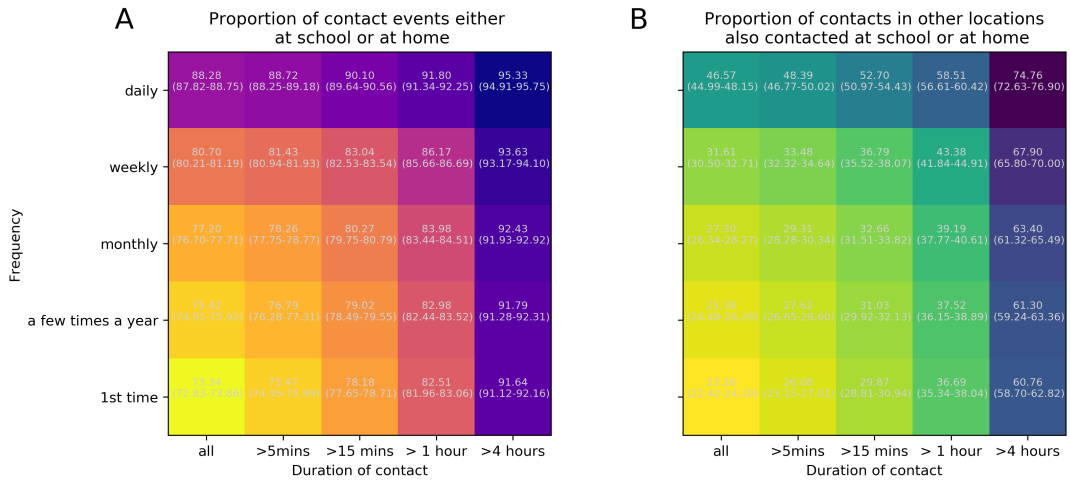


Figure 4.1 Contacts at home and at school

A) Proportion of contact events either at school or at home cumulative by frequency and duration (95% CI). B) Proportion of contacts in other locations (not school or home) also contacted at school or at home cumulative by frequency and duration (95% CI).

4.4 Discussion

Contacts between children are widely acknowledged as being of great importance in the transmission of many common infections including ‘childhood infections’ and Influenza. Within-school contact networks and household transmission have been the focus of much research in recent years [9, 10, 13–18] due to the large proportion of children’s time spent in these contexts. To establish the importance of school and household contacts for transmission between children I have analysed existing social contact data.

My analysis suggested that children have more contacts outside the household than other age groups. The majority of these contacts occur in school. They are also more likely to

involve touch than contacts between adults, which might indicate higher risk of transmission on contact.

A larger majority of contacts were made in school and home when only long duration and contacts that were made more frequently were considered. This showed that the dominance of school and home settings for contact between children increased for more established contact relationships, which occur over long durations, suggesting that contact events outside of these contexts are more likely to be 'incidental' short and events rarely repeated.

I estimated the proportion of contact time between children to be over 80% in school or households overall and nearly 90% in schools or households on weekdays.

This analysis is intended to show in principle, that a large proportion of contacts amongst school-aged children occur at school. There are some severe limitations to the approach, which largely stem from the data on which it is based.

Contacts were self-reported, which leaves room for interpretation from participants. Biases could be introduced if some participants, for example, report large groups with whom they met or just a small number of the group with whom they had the most contact. In particular, children may include whole classes, therefore over-reporting contacts at school, however there is evidence that fewer school contacts are reported through surveys than observed by sensor-based studies [20]. The survey also relies on the participants' memory, so there may also be a tendency to report contacts which are easier to remember. This effect could be amplified since contacts were right-censored to 30 contacts per

participant, due to the paper-based format of the survey [19]. This may have meant that some contacts could not be reported, especially for children, who frequently report large numbers contacts. The way this was dealt with by the participants may vary and may impact the results. For example, children may have systematically favoured reporting school and home contacts over those in other places because they were easier to recall.

The duration of contacts was only reported per contact, not per contact event, so I had to estimate the distribution of time between contacts where multiple events in different locations were reported. This however accounted for only 10% of contacts. More importantly, durations were not reported as an exact time, but instead as a band. To give a specific value to this I estimated the time as the midpoint of the band. This assumption is not tested and cannot be expected to necessarily result in accurate estimates of duration of contact when aggregated. For this reason, the reported values of the proportion of contact time in schools and homes should not be interpreted as accurate measurements but rather to give a broad indication of how child-to-child contact time is distributed. I judge the approach to be appropriate for such broad interpretation.

If longer contact with an infected individual constitutes greater exposure to infection, the analysis indicates that, for school-aged children, the vast majority of exposure to infection occurs in schools and households. Furthermore, if ‘closer’ contact occurs between familiar individuals, made more frequently and for longer duration, it appears that the majority of such contacts also occur at school and home.

4.5 References

1. Baguelin M, Flasche S, Camacho A, Demiris N, Miller E, Edmunds WJ. **Assessing Optimal Target Populations for Influenza Vaccination Programmes: An Evidence Synthesis and Modelling Study**. *PLoS Med*. 2013, 10:e1001527. doi:10.1371/journal.pmed.1001527.

2. Worby CJ, Chaves SS, Wallinga J, Lipsitch M, Finelli L, Goldstein E. **On the relative role of different age groups in influenza epidemics.** *Epidemics*. 2015, 13:10–6. doi:10.1016/j.epidem.2015.04.003.

3. Jackson C, Mangtani P, Fine P, Vynnycky E. **The effects of school holidays on transmission of varicella zoster virus, England and Wales, 1967-2008.** *PLoS One*. 2014, 9:e99762. doi:10.1371/journal.pone.0099762.

4. Jackson C, Mangtani P, Vynnycky E, Fielding K, Kitching A, Mohamed H, et al. **School closures and student contact patterns.** *Emerg Infect Dis*. 2011, 17:245–7. doi:10.3201/eid1702.100458.

5. Te Beest DE, Birrell PJ, Wallinga J, De Angelis D, Van Boven M. **Joint modelling of serological and hospitalization data reveals that high levels of pre-existing immunity and school holidays shaped the influenza A pandemic of 2009 in The Netherlands**. *J R Soc Interface*. 2014, 12. doi:10.1098/rsif.2014.1244.

6. Melegaro A, Jit M, Gay N, Zagheni E, Edmunds WJ. **What types of contacts are important for the spread of infections?: using contact survey data to explore European mixing patterns.** *Epidemics*. 2011, 3:143–51. doi:10.1016/j.epidem.2011.04.001.

7. Kucharski AJ, Wenham C, Brownlee P, Racon L, Widmer N, Eames KTD, et al. **Structure and consistency of self-reported social contact networks in British secondary schools**. 2018. doi:10.1371/journal.pone.0200090.

8. Guclu H, Read J, Vukotich CJ, Galloway DD, Gao H, Rainey JJ, et al. **Social Contact Networks and Mixing among Students in K-12 Schools in Pittsburgh, PA**. 2016. doi:10.1371/journal.pone.0151139.

9. House T, Baguelin M, Van Hoek AJ, White PJ, Sadique Z, Eames K, et al. **Modelling the impact of local reactive school closures on critical care provision during an influenza pandemic**. *Proc R Soc B Biol Sci*. 2011, 278:2753–60. doi:10.1098/rspb.2010.2688.

10. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton J-F, et al. **High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School**. *PLoS One*. 2011, 6:e23176. doi:10.1371/journal.pone.0023176.

11. Luh D-L, You Z-S, Chen S-C. **Comparison of the social contact patterns among school-age children in specific seasons, locations, and times.** *Epidemics*. 2016, 14:36–44. doi:10.1016/j.epidem.2015.09.002.

12. Eames KTD, Tilston NL, Edmunds WJ. **The impact of school holidays on the social mixing patterns of school children**. *Epidemics*. 2011, 3:103–8. doi:10.1016/j.epidem.2011.03.003.

13. Salathe M, Kazandjieva M, Lee JW, Levis P, Feldman MW, Jones JH. **A high-resolution human contact network for infectious disease transmission**. *Proc Natl Acad Sci*. 2010, 107:22020–5. doi:10.1073/pnas.1009094108.

14. Fournet J, Barrat A. **Contact Patterns among High School Students**. *PLoS One*. 2014, 9:e107878. doi:10.1371/journal.pone.0107878.

15. Kucharski AJ, Conlan AJK, Eames KTD. **School's Out: Seasonal Variation in the Movement**

Patterns of School Children. *PLoS One*. 2015, 10:e0128070. doi:10.1371/journal.pone.0128070.

16. Potter GE, Handcock MS, Longini IM, Halloran ME. **Estimating within-school contact networks to understand influenza transmission**. *Ann Appl Stat*. 2012, 6:1–26. doi:10.1214/11-AOAS505.

17. Conlan AJK, Eames KTD, Gage JA, von Kirchbach JC, Ross J V, Saenz RA, et al. **Measuring social networks in British primary schools through scientific engagement.** *Proc Biol Sci*. 2011, 278:1467–75. doi:10.1098/rspb.2010.1807.

18. Hens N, Ayele GM, Goeyvaerts N, Aerts M, Mossong J, Edmunds JW, et al. **Estimating the impact of school closure on social mixing behaviour and the transmission of close contact infections in eight European countries.** *BMC Infect Dis*. 2009, 9:187. doi:10.1186/1471-2334-9-187.

19. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. **Social contacts and mixing patterns relevant to the spread of infectious diseases.** *PLoS Med*. 2008, 5:e74. doi:10.1371/journal.pmed.0050074.

20. Grantz, H. K, Cummings DAT, Zimmer SM, Vukotich CJ, Galloway DD, Schweizer M Lou, et al. **Age-specific social mixing of school-aged children in a US setting using proximity detecting sensors and contact surveys**. 2020. doi:10.1101/2020.07.12.20151696.

5 Two frameworks for analysing social structure and disease transmission using national school data.

Objective: *Develop a framework to quantify social structure within contact networks of school children in a way that can be used in an infectious disease transmission model.*

5.1 Introduction

The purpose of this chapter is to introduce to two frameworks that I have developed to make use of national school data for analysing social structure and the implications for transmission of infectious disease in school-aged children. These frameworks are used in later chapters to analyse how social structure impacts transmission dynamics.

The role of social structure in disease transmission

Transmission of directly transmitted pathogens requires contact between a susceptible and infected member of the host population. This places 'contact events' at the centre of transmission dynamics[1]. A contact event describes an instance where two potential hosts have an encounter that could result in transmission if one is infected with and the other susceptible to infection with a particular pathogen. Many epidemiological phenomena are thought to depend on the setting, duration and frequency of contact events[2–6]. Much effort has been spent on studying the measured or inferred properties of these contact events over several spatial and temporal scales [2, 3, 7–17].

Studying contact networks in a population of humans

Broadly, there are two groups of methods for studying contact networks: measurement by collection of new data, and inference by use of existing data.

Firstly, direct measurement of contacts. Social contact can be measured directly, most frequently at an individual level. A growing number of contact surveys have been undertaken to elicit a generalised understanding of how individuals interact with the

population around them. Such studies have been instrumental in our understanding of transmission between age groups, distance between residence and contact sites, and the locations and activities during which contact most frequently occurs[2, 8, 9, 16, 18–22]. In addition, there have been a number of studies, which utilize location-tracking devices to measure movement of individuals. By providing such devices to every member of the a population, contact events can be registered when participants remain in close proximity for a threshold period of time [13, 23, 24].

Secondly, Existing data can often be useful to infer long-range contact between populations for example between cities[7], nations[25] or contained populations. For example, long distance contact within a population of farm animals can be inferred from transport networks between farms, where infected livestock might be traded and introduce infection into new herds[26]. Similarly, airline passenger data has been successful in predicting the arrival time of new strains of influenza across the world[27]. More recently, the transmission of healthcare related infections between institutions have been studied by use of patient transfer data[28–31].

Within geographic areas with relatively continuous population densities, transmission risk has often been assumed to decay with a power-law relationship to distance from an infected host, chiefly driven by observed movements of people following this relationship [32]. This is also supported by evidence from social contact surveys carried out in multiple settings. This assumption has been a sufficient approximation in some cases, however there is also evidence that it falls short of properly accounting for the impact of spatial considerations on an epidemic[33]. Of particular interest to this analysis, where the population is structured into highly segregated social groups, such approximations are

unable to capture the particularities of transmission between them. Social groups are characterised by collections of individuals who possess a commonality with which they choose to identify, and who show strong preference for relationship (or contact) within the group relative to outside the group. Such behaviour has been observed in sub-populations of particular ethnic, socio-economic and religious and political descriptions. Complex social structures mean that heterogeneity in transmission and uptake of interventions, such as vaccination, correlated with particular social groups may be important for transmission dynamics and the effectiveness of healthcare interventions, especially if the pathogen is close to elimination.

A justification for using school data

Approaching social structure from an individual basis, e.g. through social contact surveys, is challenging for multiple reasons. Alongside standard information about the contact (for example: age, sex, duration and frequency of contact) the additional information required in a survey to identify clustering of contacts within social groups would require a substantial increase in burden on the survey participants. This is likely to reduce the quality of responses and response rate[34, 35]. Moreover, it is not always immediately clear if contacts are part of a particular group, making accurate and consistent recording of contacts challenging for the participants. Finally, by increasing the stratification of contacts to include ethnic or social group the required sample size would increase greatly. The combination of these factors makes survey methodology inappropriate within the scope of my current work. In lieu of a dedicated social contact survey, certain government-collected data may provide information that is useful for inferring properties of social structure.

Children dominate transmission in many instances as observed in Influenza A H1N1 outbreaks in London and Birmingham (studied in Chapter 3). This is also true for seasonal influenza[36] and many childhood infections such as measles, mumps and rubella. This suggests that by understanding the social structure of childhood contacts (between school-aged children), insight into how social structure impacts transmission of a number of important infections might be gained. Furthermore, there is evidence that the vast majority of infectious contact between children of school-going age occurs in the home and at school (chapter 4), there is also evidence that these contacts are of particular importance for transmission of particular pathogens circulating in children in Europe [5].

Many governments routinely collect data from schools to support service assessment and policy making[37]. This data contains information about pupils who attend each school including residence address. British and Dutch governments both publish aggregates of this data [38] and invite applications for bespoke data from their databases. National school data linked to socio-demographic data from the national census could provide important source of insight into how social, ethnic and religious groups integrate through the school system. In addition, by analysing the differences in the school system in different areas, heterogeneity in the epidemiology of outbreaks may be explained.

In this chapter I present how national school data can be used to evaluate interaction of school-aged children from different social groups and construct models of infectious transmission. These frameworks can be used to provide insight into the impact of social structure on transmission dynamics.

5.2 Proposed Frameworks

Analysis framework summary

I propose two frameworks, which use national school data to investigate the role of the school system in determining the epidemiology of an outbreak:

The first framework provides a means to evaluate preferential mixing within ethnic and socio-economic groups within schools based on residence data of pupils. I use this to assess clustering of social groups within London.

The second framework provides a means to construct a network of schools where the links between the schools give the strength of contact between them. I used this framework in the analyses detailed in the following chapters of this thesis to evaluate transmission within school-age populations in two distinct settings (Analyses C and D - chapters 6 to 8).

Framework 1: Integration of ethnic and social groups through schools

The aim of the first framework is to measure the relative rate of contact between social groups (such as ethnicities or socio-economic groups) through the school system. This framework requires data containing pupils' residence aggregated at a small geographical area, and the social and ethnic breakdown for the same geographical areas at the same resolution as the pupil residence data, for example national census data. From these data an estimate can be made for the relative rate of contact between children within and between social and ethnic groups at school.

First, I constructed two matrices using schools' data. The first gives the proportion of children in each geographical area that attend each school ($\mathbf{a}_{ls}$), the second the proportion of children in each school who reside in each geographical area ($\mathbf{s}_{sl}$). Assuming proportional mixing in schools, a matrix of the interaction between geographical areas, through schools, is calculated as the product of $\mathbf{a}_{ls}$ and $\mathbf{s}_{sl}^{\mathbf{T}}$:

$$\mathbf{L} = \mathbf{a}_{ls} \cdot \mathbf{s}_{sl}^{\mathbf{T}}$$

Secondly, using national census data, I constructed two vectors, which capture the distribution of social-groups in the population. One vector, e_a, gives the proportion of each geographical area who are in social group A (i.e. each element gives the proportion of social group A in the total population who reside in a particular area), the second vector, ϵ_b, gives the proportion of social group B who are in each geographical area (i.e. each element gives the proportion of the total population of social group B who reside in a particular area).

Using, e_a, and the interaction between areas, **L**, I calculated a vector of the proportion of contacts in each area that belong to social group A.

$$\mathrm{c}_a = \mathbf{L} \cdot \mathrm{e}_a$$

Finally, using ϵ_b I calculated the proportion of social group B's contacts that belong to social group A.

$$\mathrm{p}_{ba} = \mathrm{c}_a \cdot \epsilon_b$$

This framework relies on two clear assumptions. Firstly, I assume that children from a particular area are distributed between schools independent of socio-economic status or ethnic group. Secondly, I assume that all contact behaviour within school is homogenous with no additional preference for particular social groups to mix with each other relative to others. There is a chance that certain social groups have preference for particular schools. Also, there is evidence of positive homophily within schools (preference for mixing within social groups), particularly amongst ethnic groups[39]. Therefore, the proportion of contacts between social groups is likely to be over-estimated and the proportion of contacts within the same social group is likely to be under-estimated.

Framework 2: Connecting schools through households

Outbreaks within school populations and transmission within households are a key factor in the epidemiology of infectious disease, particularly influenza and other infections strongly associated with children. However, on their own each of these phenomena are naturally limited by the size of the school population or household. Large epidemics rely on both to spread across large populations of multiple schools and households.

I propose a second use of national school data to quantify connections between schools through household-based contacts. The aim of this framework is to provide a method for approximating how school-level outbreaks may translate to large epidemics and how the particular structure of the school network might determine the epidemiology of outbreaks in school-aged populations.

I used national school data to construct 'real networks' of schools, connected by residences in which children attend two or more institutions between them (figure 5.1 A). I weighted the edges of the network by the number of unique pairs of pupils, which form a link between two schools by residence at the same address. Where this data was not directly accessible, I weighted the edges by the number of pupils that transfer from each primary school to each secondary school each year (feeding rate). I developed a method of inferring an estimate of the number of unique contacts between pupils from feeding rate and local household size information to allow between school transmission to be estimated in settings where there is insufficient data to calculate the number of unique contact pairs directly.

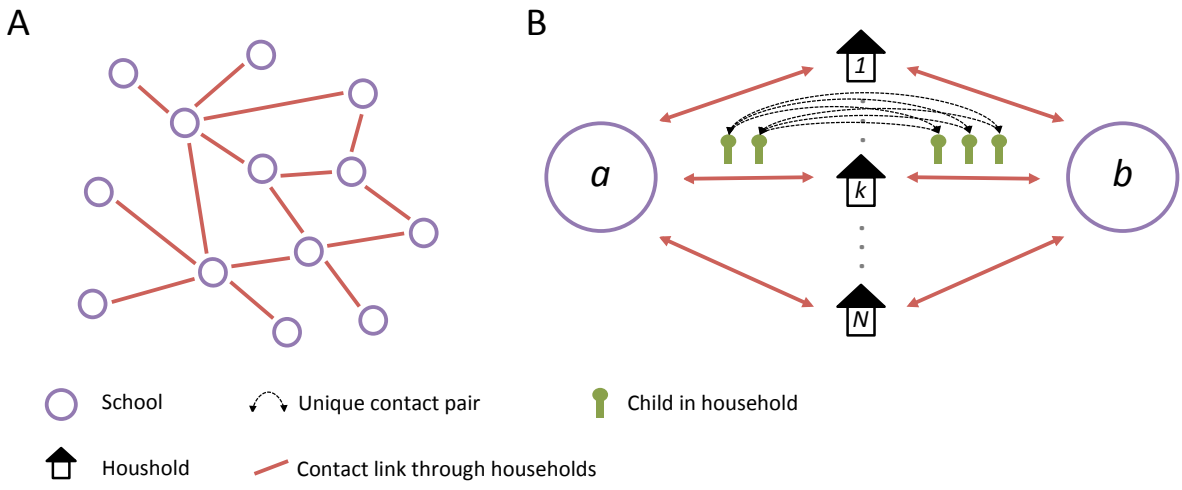


Figure 5.1 A network of schools linked by households

A) A network of schools constructed such that schools are connected when contact is made between pupils of different schools within a household. B) The strength of contact between schools is quantified by calculating the number of unique contact pairs (one child in each school). The number of pairs per household is the product of the number of children who attend school *a,* and the number of children who attend school *b*. The total number of unique pairs is the sum of unique pairs in each household with children attend both school *a* and *b*.

Where sufficient data do exist, the number of unique pairs which connect school a and b within a particular house, h_k, is given by is given by $n_{a,h_k}\, n_{b,h_k}$. Where there are n_{a,h_k} children who attend school a and n_{b,h_k} children who attend school b. For a given school with set of N shared households $H = \{h_1 : h_N\}$, the total number of contact pairs between the school is $C_{ij} = \sum_{k=1}^{N} n_{a,h_k}\, n_{b,h_k}$ (figure 5.1 B).

Construction of a network from primary to secondary feeder rate

In some settings the school data that was made available to me was not sufficient to directly calculate the number of unique contact pairs between schools. In lieu of sufficient data I developed a method of estimating the number of unique contacts between primary and secondary schools based on the rate of transition of pupils from each primary school to each secondary school following the completion of their primary education (feeder rate).

I assumed that households consist of one or multiple siblings of different ages living in the same home. In households where one or more siblings attend primary school and one or more attend secondary school, this would constitute a 'link household'. I assume that each child observed in the feeder rate data must belong to a household, which either does or doesn't form a link between the primary school and secondary school they have moved from and to. To infer the proportion of households that do form a link I first identify that the observed movement of a child from primary to secondary school must represent one of 4 possible events depending on the nature of the child's household and the age of the child relative to their siblings (figure 5.2), either:

1. The child has no siblings therefore no prior link exists through the child's household and *no link is formed* by siblings remaining in primary school after the child's transition.
2. The child has only younger siblings therefore no prior link exists, but *a link is formed* because the younger siblings remain in primary school after the child's transition.
3. The child has only older siblings therefore a prior link exists, but *this link is destroyed* because no siblings remain in primary school after the child's transition.
4. The child has both older and younger siblings therefore a prior *link exists and is maintained (not destroyed)* because some siblings remain in primary school after the child's transition.

I used the data in combination with areal household composition data to infer the number of contacts between each primary and secondary school via households.

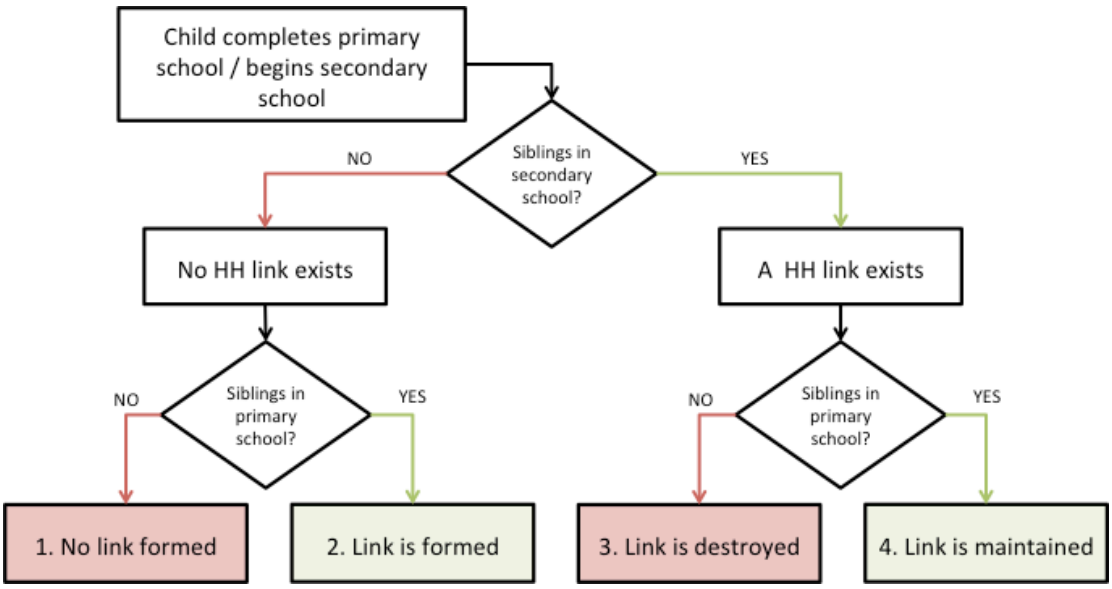


Figure 5.2 A decision tree showing the 4 potential implications of a child moving from primary to secondary school

A number of children are observed to transfer each year from each primary to each secondary school, N_{obs}^{ij}. For each primary and secondary school pair I calculated the number of households that have children in both schools for households with *n* children in the household, for each value of *n*. I defined the proportion of all children in the local population that are part of a household of *n* children to be w_n. Therefore, the number of children observed in the feeder data, who belong to a household with a total of *n* children, $N_{n,obs}^{ij}$, is given by:

$$N_{n,obs}^{ij} = w_n N_{obs}^{ij}$$

For households with multiple children, the oldest child forms a link between schools (Action 2). The following $n-2$ children neither create nor destroy the link (Action 4). The youngest child, with no younger siblings in primary school, destroys the link. Therefore, the proportion of children in a household of n children, where $n > 1$, that either form or maintain a link between the schools is given by $\frac{(n-1)}{n}$. For households with one child the same is true since the child doesn't form a link (Action 1.). The proportion of households with $n = 1$, which constitutes an ongoing link, can also be expressed as $\frac{(n-1)}{n}$.

Hence, the number of children who belong to a household with n children, that forms a link between the schools is given by:

$$\widetilde{N}_{n,obs}^{ij} = \frac{N_{n,obs}^{ij}(n-1)}{n},$$

For any value of n.

As many consecutive siblings are more than a year apart in age, each link household that exists between schools is not observed every year. I accounted for this by multiplying the subset that was observed by a correction factor. If a_{gap} is the average age gap between consecutive siblings, the rate at which each household with a link has a child completing primary school is $\frac{1}{a_{gap}}$. To find the total number of live link households with n children, I multiplied the number I observed, $\widetilde{N}_{n,obs}^{ij}$, by the average age gap between siblings, a_{gap}.

$$\widetilde{N}_{n,tot}^{ij} = a_{gap}\widetilde{N}_{n,obs}^{ij}$$

Next I calculated the average number of pairs of contacts between schools within each household of n children over the total life of the link household. If n_p is the number of children in primary and n_s is the number of children in secondary school, for a link household of n children there are n-1 combinations of n_s and n_p. Assuming the total age span of siblings is does not allow older siblings to leave secondary school before the youngest sibling joins, $n_p = (n - n_s)$. For any given configuration, the number of unique pairs between the schools is then given by $n_s(n - n_s)$. Assuming the age gap between siblings is constant, the time for which each configuration is the same. Therefore, the

average number of contact pairs formed by a link household with n children, ρ_n, is simply the mean number of unique contact pairs over all possible configurations (figure 5.3):

$$\rho_n = \frac{1}{n-1}\sum_{n_s=1}^{n-1} n_s(n-n_s)$$

The total number of contact pairs between schools, within all households of n children is therefore given by:

$$C_{ij}^{n} = \rho_n \widetilde{N}_{n,tot}^{ij}$$

To find the estimated total number of contacts between school i and school j of all household sizes, I took the sum of C_{ij}^{n} over all appropriate values of n:

$$C_{ij} = \sum_{n} C_{ij}^{n} = N_{obs}^{ij} \cdot a_{gap} \cdot \sum_{n}\left[\frac{w_n(n-1)}{n}\cdot \rho_n\right]$$

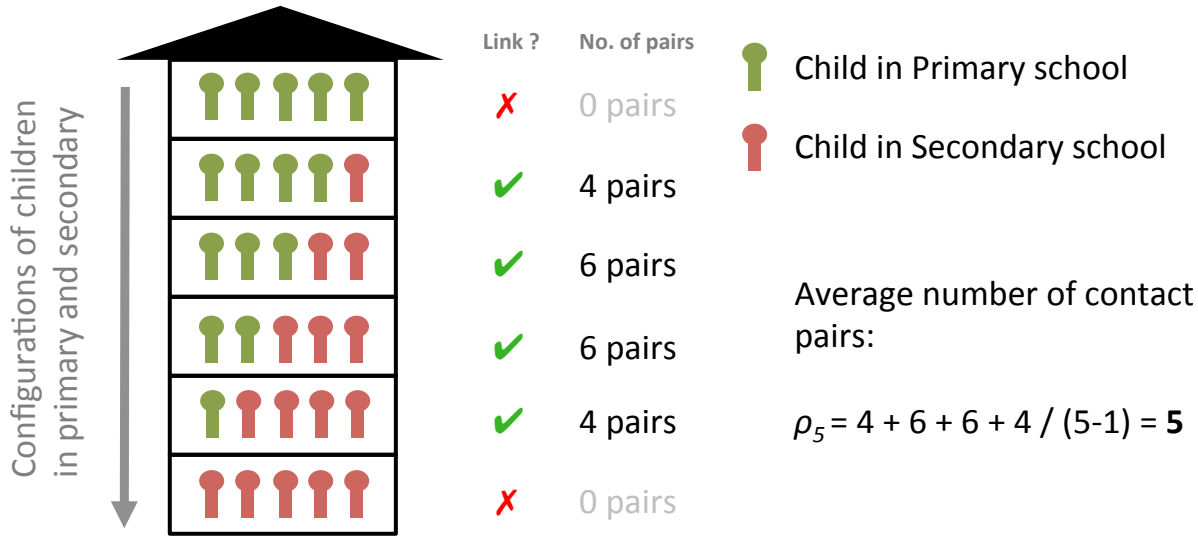


Figure 5.3 Calculating the unique number of contact pairs per household

An example of calculating the average number of unique contact pairs over the lifetime of a link household with 5 school age-children; there are 6 possible configurations of $\boldsymbol{n_p}$ children in primary school and $\boldsymbol{n_s}$ children in secondary school. Of those 6, 4 constitute a link between primary and secondary through the household. The average number of unique contact links between primary and secondary over during the time that the household provides a link is 5.

Translation of contact network to a school-based transmission model

Now that the network of contact between schools has been defined, it is used to estimate the probability that individual school outbreaks can seed an outbreak in each neighbouring school.

For each pair of neighbouring schools, I calculate the probability that an outbreak could be seeded in school i given that an outbreak does occur in adjacent school j. First, I consider the probability of transmission between siblings, in the event that one is infectious and the other is susceptible, to be a set value q. The probability that the child from school j is infected is denoted by P_j^I and the probability that the child from school j is susceptible by P_i^S. The probability that a single infected student in school i causes a large outbreak in that school is P_i^{OB}. The probability of an outbreak in school j leading to an outbreak in school i through each unique contact pair that link schools i and j is:

$$P_j^I P_i^S q P_i^{OB}$$

The probability that the child in school j is infected, P_j^I, was assumed to be equal to the proportion of the school children infected by the outbreak in that school. I assumed that this is defined by the solution of the final size equation[40]:

$$R_j(\infty) = (1 - V_j)(1 - e^{-(1-V_j)R_0 R(\infty)})$$

Where V_j is the vaccination coverage in school j.

The probability that the child in school i is susceptible, P_i^S is assumed to be equal to the proportion of school i that remains unvaccinated, $(1 - V_i)$.

The probability of an outbreak occurring in that school as a result of this transmission can be written in terms of the effective reproduction number, R_{eff}:

$$P_i^{OB} = \left(1 - \frac{1}{R_{eff}}\right) = \left(1 - \frac{1}{(1 - V_i)R_0}\right)$$

which assumes a geometric distribution of within-school contact rate amongst children[40].

The probability that none of the unique contact pairs causes an outbreak in school i can be written:

$$\prod_{All\ pairs} \left(1 - P_j^I P_i^S q P_i^{OB}\right) = \left(1 - P_j^I P_i^S q P_i^{OB}\right)^{C_{ij}}$$

Therefore, the probability that at least one contact pair causes an outbreak in school i is the complement of this:

$$P_{trans,ij} = \left[1 - \left(1 - P_j^I P_i^S q P_i^{OB}\right)^{C_{ij}}\right]$$

This provides a basis upon which to model simulations of outbreaks across networks of schools in different settings. And is used later in this thesis for analysis of influenza and

social groups in London (chapter 6) and measles outbreaks in the Netherlands (chapter 7 and 8).

5.3 Methods

To evaluate interaction between social groups in the school-aged population in London, I used Framework 1 (detailed above) to estimate the baseline rate of contact between socio-economic and ethnic groups. I estimated the rate of contact between ethnic groups and deprivation deciles relative to what would be expected through proportional mixing for the whole population of London. Proportional mixing was taken as for example: if an ethnic group comprised 40% of the population that ethnic group would be expected to comprise 40% of each child's contacts. I repeated the same analysis for deprivation deciles.

The matrices $\mathbf{a}_{ls}$ and $\mathbf{s}_{sl}$ (to determine interaction between schools are geographic areas) were constructed from data accessed from the London data service Schools Atlas[38], which is a service provided by the Greater London Authority (GLA). The data included the number of children in each school (n. 2838) who reside in each Lower Super output Area (LSOA) (n. 5780) and vice versa for all LSOAs and for every state funded school in the Greater London Authority area. The data was constructed by GLA by aggregating pupil level data from the 2016 Autumn schools census, which is submitted to the Department for Education by each school individually, in October, January and May each year. The data only includes state funded institutions providing primary education (4 – 11 year olds), secondary education (11 – 16 year olds) and further education (typically

16 – 18 year olds). Independent schools are not included, which amount to 130 schools in the GLA area.

The data for vectors e_a and ϵ_a, both regarding ethnic group and deprivation decile, were calculated using United Kingdom Census data from 2011[41]

To assess how preferential mixing within social groups breaks down over multiple generations of contact (e.g. 2 generations of contact reflects contacts of contacts), I estimated the proportion of the nth generation of contacts in each LSOA from each LSOA as $\mathbf{L}^{\mathrm{n}}$. Hence the proportion of nth generation contacts of ethnic group b that are ethnic group a is given by:

$$p_{ba,gen=n} = (\mathbf{L}^{\mathrm{n}} \cdot \mathrm{e}_a) \cdot \epsilon_{\mathrm{b}}$$

I calculated the value for the within ethnic group proportion of nth generation contacts, $p_{ba=b,gen=n}$, for each ethnic group for generations up to 200.

5.4 **Results**

In general, ethnic groups mixed within their own group more than expected. The ethnic group that mixed most disproportionately within their own group were Bangladeshi, whose contacts were 6.7 times more likely to be Bangladeshi than would be expected by proportionate mixing (Figure 5.4). Broader ethnic groups appeared to mix with each other more frequently also, particularly Asians, where Indians, Pakistanis and Bangladeshis all mixed with each other more than expected. This was also true of black ethnicities

(Caribbean, African and Other Black) although this was to a lesser extent than Asian ethnicities.

The same was true when stratifying the population by deprivation status. Children were more likely than expected to mix with others of the same or similar deprivation status (Figure 5.5). This was particularly true of the most affluent and most deprived deciles, where children mixed with others in their own deprivation decile 6.1 and 9.3 times more than expected respectively.

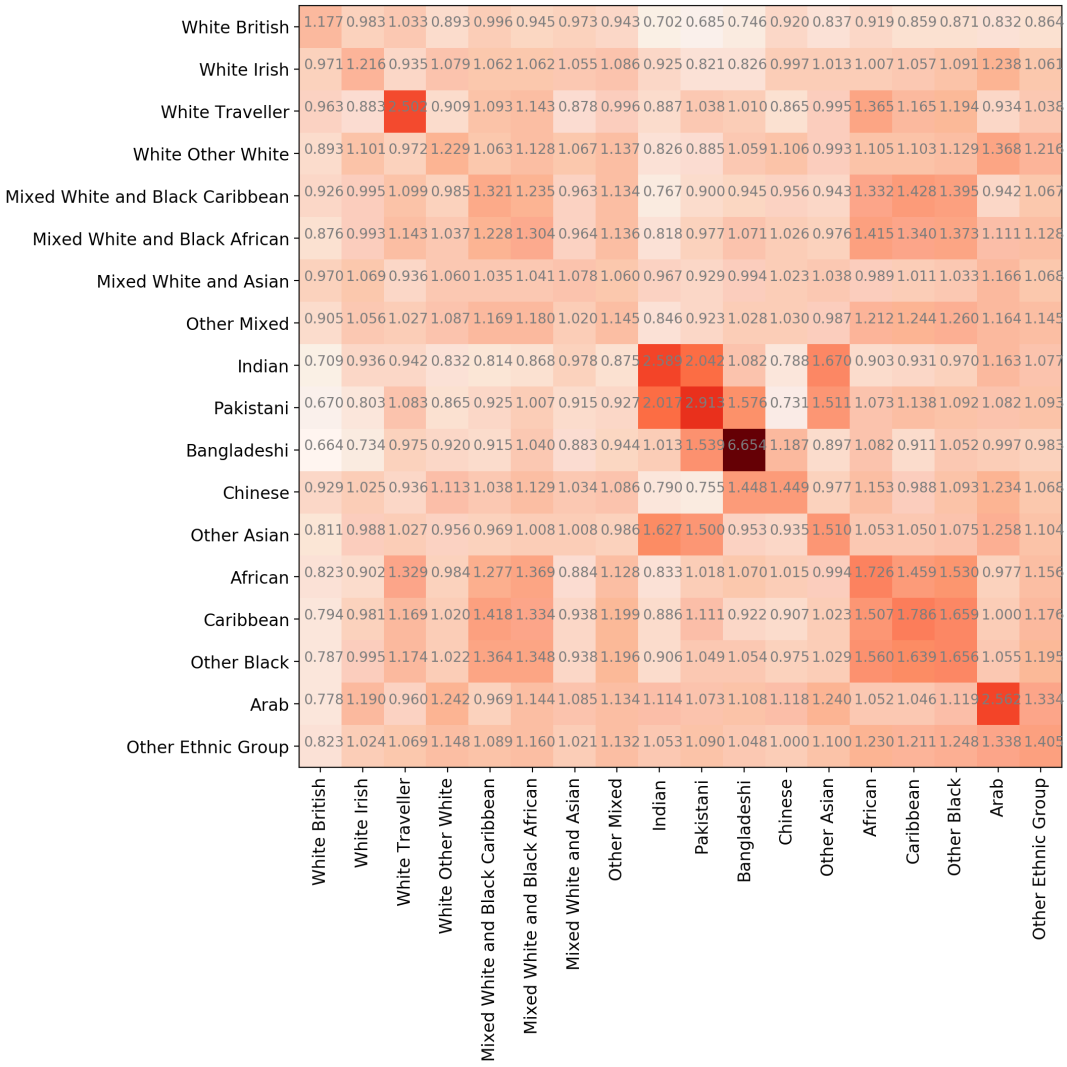


Figure 5.4 Proportion of contacts in each ethnic group by ethnic group, relative to proportion of the population.

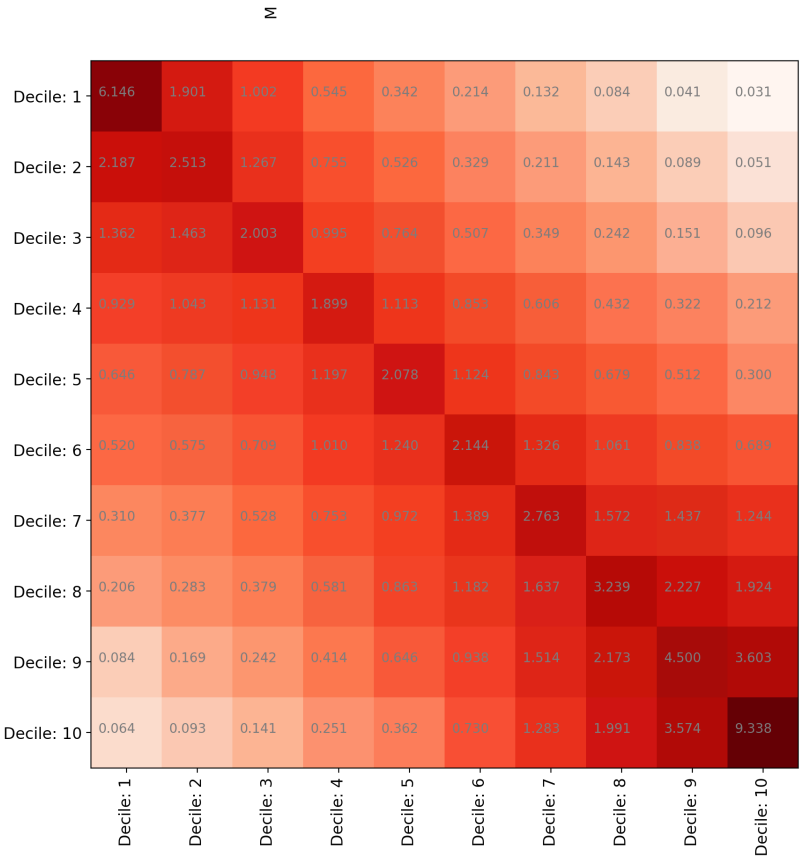


Figure 5.5 Proportion of contacts in each deprivation decile by deprivation decile, relative to proportion of the population.

Over multiple generations of contact, disproportionate representation of contacts within each ethnic group reduces. However, at 200 generations of contacts there is still substantially disproportionately high representation from the same ethnic group for Asians. This was particularly true for Bangladeshi children, where there was still 2.5 times as many 200th generation contacts that were Bangladeshi than would be expected by proportional mixing. Similarly, integration of deprivation deciles improves at increasing generations of contact.

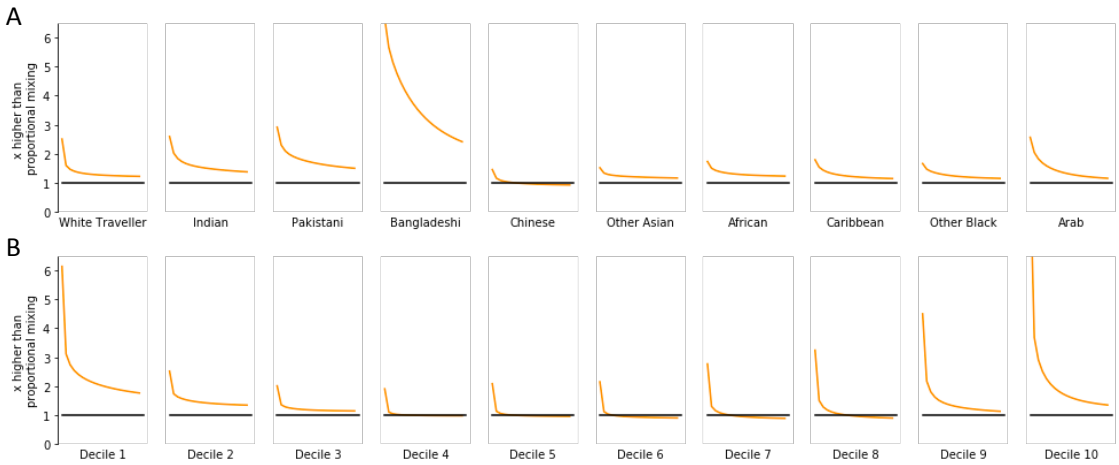


Figure 5.6 Proportion of contacts in the same social group after n generations of contacts

A) Proportion of contacts in the same ethnic group by ethnic group over the first 200 generations of contacts B) Proportion of contacts in the same deprivation decile by deprivation decile over the first 200 generations of contacts

5.5 **Discussion of results**

Quantifying the extent of interaction between social groups is a key part of understanding how different populations' experience of an outbreak of infectious disease may differ, leading to measured inequalities in risk. I have used national school data to assess the opportunities for children from different ethnic groups and deprivation quintiles to interact through schools.

In general, children mixed more than expected within their own ethnic group and deprivation status than would be expected if mixing were proportional to ethnic group population size. The most affluent and most deprived areas appear to be most likely to mix more within their own group. Likewise, Bangladeshi children mixed most disproportionately within their own ethnic group.

At multiple generations of contacts, for the majority of ethnic groups, representation of ethnicity amongst contacts becomes proportionate to the population within the first 20 generations. For Bangladeshi children, however, the 200th generation contacts were still 2.5 times more likely to be Bangladeshi than expected by proportional mixing, which suggests that Bangladeshi children are relatively clustered within the school system, even over multiple generations of contact. Similarly, for deprivation status, increasing generations of contacts generally moved out of the similar deprivation status groups.

The effects observed suggest that early phases of an outbreak may stay within particular ethnic or social groups, in the case where transmission between school children is driving the dynamics of the outbreak. However, the assimilation to proportionate representation over the course of several generations of contacts suggests that the outbreak would become well-mixed within a few generations of transmission. The high level of over representation of Bangladeshi's both in first degree contacts and slower assimilation to proportionate representation of Bangladeshi contacts indicates that the effects are likely to be most pronounced for this group than others, and it may take longer for an outbreak to become more well mixed in the population.

This analysis makes some major assumptions about school attendance and contact behaviour within schools. Firstly, the framework assumes that school choice of children from a particular LSOA is independent of ethnicity. This assumption is necessary for this analysis, as no clear data on trends in school choice by ethnicity are available at this time. It is likely however that certain ethnic groups are more likely to choose particular schools, particularly when schools assume a particular religious identity. Secondly, this analysis assumes proportional mixing within schools. Again, there is a chance that there is some

degree of ethnic homophily within friendships formed within schools, as has been found in previous studies [39]. This could also be true for socio-economic status. Homophily of this kind, however, is unquantified for the setting assessed here, and may vary significantly between ethnic groups and schools. Both of these factors would serve to increase segregation between ethnic groups and deprivation quintiles; therefore, this analysis is likely to under-predict the overall segregation within the school system, but serves to provide a minimum necessary segregation of children through schools.

Finally, the data I used for this analysis only includes state-funded schools. Independent schools could impact the results in two ways. Firstly, since they account for the education of some proportion of the school-aged population which will provide links between LSOAs in a way that is not captured here. Secondly, although admission to independent schools is becoming more equitable, there remains an overrepresentation of white and affluent children in these schools. This may serve to further isolate children by ethnicity and deprivation status.

In conclusion, analysis of data on residence of pupils of London schools supports the notion that ethnic and social groups are necessarily segregated through the school system. This may be a contributing factor in observed inequalities in risk between ethnic groups and areas of differing deprivation status. This is especially likely in the early phase of an outbreak but might be expected to reduce over a number of generations of transmission. Outbreaks that initiate in Asian communities or highly affluent or highly deprived areas are most likely to lead to observed disparities. Higher risk in Bangladeshi groups may last for more of the outbreak as disproportionate representation of contacts remains for a high number of generations.

5.6 Summary

This chapter introduces two frameworks for analysing national school data for the purpose of inferring relative contact between social groups and identifying routes through which infectious pathogens are likely to spread through the population. In the remaining analysis chapters, I have applied these methods to research questions around the relative risk of infection within different social groups in London and the role schools might play in clustering of unvaccinated children in the Netherlands.

5.7 References

1. Edmunds WJ, O'Callaghan CJ, Nokes DJ. **Who mixes with whom? A method to determine the contact patterns of adults that may lead to the spread of airborne infections.** *Proc Biol Sci*. 1997, 264:949–57. doi:10.1098/rspb.1997.0131.

2. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. **Social contacts and mixing patterns relevant to the spread of infectious diseases.** *PLoS Med*. 2008, 5:e74. doi:10.1371/journal.pmed.0050074.

3. Danon L, Read JM, House TA, Vernon MC, Keeling MJ. **Social encounter networks: characterizing Great Britain.** *Proc Biol Sci*. 2013, 280:20131037. doi:10.1098/rspb.2013.1037.

4. De Cao E, Zagheni E, Manfredi P, Melegaro A. **The relative importance of frequency of contacts and duration of exposure for the spread of directly transmitted infections.** *Biostatistics*. 2014, 15:470–83. doi:10.1093/biostatistics/kxu008.

5. Melegaro A, Jit M, Gay N, Zagheni E, Edmunds WJ. **What types of contacts are important for the spread of infections?: using contact survey data to explore European mixing patterns.** *Epidemics*. 2011, 3:143–51.

doi:10.1016/j.epidem.2011.04.001.

6. Read JM, Eames KTD, Edmunds WJ. **Dynamic social networks and the implications for the spread of infectious disease.** *J R Soc Interface*. 2008, 5:1001–7. doi:10.1098/rsif.2008.0013.

7. Viboud C, Bjornstad ON, Smith DL, Simonsen L, Miller MA, Grenfell BT. **Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza**. *Science (80-)*. 2006, 312:447–51. doi:10.1126/science.1125237.

8. Stein ML, van Steenbergen JE, Chanyasanha C, Tipayamongkholgul M, Buskens V, van der Heijden PGM, et al. **Online respondent-driven sampling for studying contact patterns relevant for the spread of close-contact pathogens: a pilot study in Thailand.** *PLoS One*. 2014, 9:e85256. doi:10.1371/journal.pone.0085256.

9. Stein ML, van der Heijden PGM, Buskens V, van Steenbergen JE, Bengtsson L, Koppeschaar CE, et al. **Tracking social contact networks with online respondent-driven detection: who recruits whom?**. *BMC Infect Dis*. 2015, 15:522. doi:10.1186/s12879-015-1250-z.

10. Conlan AJK, Eames KTD, Gage JA, von Kirchbach JC, Ross J V, Saenz RA, et al. **Measuring social networks in British primary schools through scientific engagement.** *Proc Biol Sci*. 2011, 278:1467–75. doi:10.1098/rspb.2010.1807.

11. Kucharski AJ, Wenham C, Brownlee P, Racon L, Widmer N, Eames KTD, et al. **Structure and consistency of self-reported social contact networks in British secondary schools**. 2018. doi:10.1371/journal.pone.0200090.

12. Guclu H, Read J, Vukotich CJ, Galloway DD, Gao H, Rainey JJ, et al. **Social Contact Networks and Mixing among Students in K-12 Schools in Pittsburgh, PA**. 2016. doi:10.1371/journal.pone.0151139.

13. Leecaster M, Toth DJA, Pettey WBP, Rainey JJ, Gao H, Uzicanin A, et al. **Estimates of Social Contact in a Middle School Based on Self-Report and Wireless Sensor Data.** *PLoS One*. 2016, 11:e0153690. doi:10.1371/journal.pone.0153690.

14. Towers S, Feng Z. **Social contact patterns and control strategies for influenza in the elderly.** *Math Biosci*. 2012, 240:241–9. doi:10.1016/j.mbs.2012.07.007.

15. Zagheni E, Billari FC, Manfredi P, Melegaro A, Mossong J, Edmunds WJ. **Using time-use data to parameterize models for the spread of close-contact infectious diseases.** *Am J Epidemiol*. 2008, 168:1082–90. doi:10.1093/aje/kwn220.

16. Danon L, House TA, Read JM, Keeling MJ. **Social encounter networks: collective properties and disease transmission**. *J R Soc Interface*. 2012, 9:2826–33.

doi:10.1098/rsif.2012.0357.

17. Read JM, Lessler J, Riley S, Wang S, Tan LJ, Kwok KO, et al. **Social mixing patterns in rural and urban areas of southern China.** *Proc Biol Sci*. 2014, 281:20140268. doi:10.1098/rspb.2014.0268.

18. Béraud G, Kazmercziak S, Beutels P, Levy-Bruhl D, Lenne X, Mielcarek N, et al. **The French Connection: The First Large Population-Based Contact Survey in France Relevant for the Spread of Infectious Diseases.** *PLoS One*. 2015, 10:e0133203. doi:10.1371/journal.pone.0133203.

19. Kiti MC, Kinyanjui TM, Koech DC, Munywoki PK, Medley GF, Nokes DJ. **Quantifying age-related rates of social contact using diaries in a rural coastal population of Kenya.** *PLoS One*. 2014, 9:e104786. doi:10.1371/journal.pone.0104786.

20. Johnstone-Robertson SP, Mark D, Morrow C, Middelkoop K, Chiswell M, Aquino LDH, et al. **Social mixing patterns within a South African township community: implications for respiratory disease transmission and control.** *Am J Epidemiol*. 2011, 174:1246–55. doi:10.1093/aje/kwr251.

21. Grijalva CG, Goeyvaerts N, Verastegui H, Edwards KM, Gil AI, Lanata CF, et al. **A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of Peru.** *PLoS One*. 2015, 10:e0118457. doi:10.1371/journal.pone.0118457.

22. Fu Y chih, Wang D-WW, Chuang J-HH. **Representative Contact Diaries for Modeling the Spread of Infectious Diseases in Taiwan**. *PLoS One*. 2012, 7:e45113. doi:10.1371/journal.pone.0045113.

23. Smieszek T, Castell S, Barrat A, Cattuto C, White PJ, Krause G. **Contact diaries versus wearable proximity sensors in measuring contact patterns at a conference: method comparison and participants' attitudes.** *BMC Infect Dis*. 2016, 16:341. doi:10.1186/s12879-016-1676-y.

24. Smieszek T, Barclay VC, Seeni I, Rainey JJ, Gao H, Uzicanin A, et al. **How should social mixing be measured: comparing web-based survey and sensor-based methods.** *BMC Infect Dis*. 2014, 14:136. doi:10.1186/1471-2334-14-136.

25. Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. **Pandemic Potential of a Strain of Influenza A (H1N1): Early Findings**. *Science (80-)*. 2009, 324:1557–61. doi:10.1126/science.1176062.

26. Woolhouse ME., Shaw D., Matthews L, Liu W-C, Mellor D., Thomas M.

Epidemiological implications of the contact network structure for cattle farms and the 20–80 rule. *Biol Lett*. 2005, 1:350–2. doi:10.1098/rsbl.2005.0331.

27. Brockmann D, Helbing D. **The Hidden Geometry of Complex, Network-Driven Contagion Phenomena**. *Science (80-)*. 2013, 342:1337–42. doi:10.1126/science.1245200.

28. Donker T, Wallinga J, Slack R, Grundmann H. **Hospital Networks and the Dispersal of Hospital- Acquired Pathogens by Patient Transfer**. 2012. doi:10.1371/journal.pone.0035002.

29. Donker T, Wallinga J, Grundmann H. **Patient Referral Patterns and the Spread of Hospital-Acquired Infections through National Health Care Networks**. *PLoS Comput Biol*. 2010, 6. doi:10.1371/journal.pcbi.1000715.

30. Donker T, Henderson KL, Hopkins KL, Dodgson AR, Thomas S, Crook DW, et al. **The relative importance of large problems far away versus small problems closer to home: insights into limiting the spread of antimicrobial resistance in England**. *BMC Med*. 2017, 15:86. doi:10.1186/s12916-017-0844-2.

31. Donker T, Smieszek T, Henderson KL, Johnson AP, Walker AS, Robotham J V. **Measuring distance through dense weighted networks: The case of hospital-associated pathogens**. doi:10.1371/journal.pcbi.1005622.

32. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel V, et al. **On the Use of Human Mobility Proxies for Modeling Epidemics**. *PLoS Comput Biol*. 2014, 10:e1003716. doi:10.1371/journal.pcbi.1003716.

33. Keeling MJ, Danon L, Vernon MC, House TA. **Individual identity and movement networks for disease metapopulations**. *Proc Natl Acad Sci*. 2010, 107:8866–70. doi:10.1073/pnas.1000416107.

34. Galea S, Tracy M. **Participation Rates in Epidemiologic Studies**. *Ann Epidemiol*. 2007, 17:643–53. doi:10.1016/j.annepidem.2007.03.013.

35. Bajardi P, Vespignani A, Funk S, Eames KT, Edmunds WJ, Turbelin C, et al. **Determinants of follow-up participation in the Internet-based European influenza surveillance platform Influenzanet.** *J Med Internet Res*. 2014, 16:e78. doi:10.2196/jmir.3010.

36. Worby CJ, Chaves SS, Wallinga J, Lipsitch M, Finelli L, Goldstein E. **On the relative role of different age groups in influenza epidemics.** *Epidemics*. 2015, 13:10–6. doi:10.1016/j.epidem.2015.04.003.

37. HM Government, Deptrtment for Education, UK. **National Pupil Database**.

https://data.gov.uk/dataset/9e0a13ef-64c4-4541-a97a-f87cc4032210/national-pupil-database.

38. Authority GL. **London Schools Atlas**. 2019. https://data.london.gov.uk/dataset/london-schools-atlas.

39. Currarini S, Jackson MO, Pin P. **An Economic Model of Friendship: Homophily, Minorities, and Segregation**. *Econometrica*. 2009, 77:1003–45. doi:10.3982/ECTA7528.

40. Diekmann O, Heesterbeek JAP. **Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation**. *Wiley Ser*. 2000, :322. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471492418.html.

41. Office for National Statistics. **http://www.neighbourhood.statistics.gov.uk/**. 2007.

6 Analysis C: Modelling influenza outbreaks on a school network in London: geographic, ethnic and socio-economic heterogeneity in risk

Objective: *Evaluate the potential role of schools in creating inequalities in influenza outbreaks in London.*

6.1 Introduction

As discussed in chapter 2, inequalities in influenza outcomes were observed between socio-economic and ethnic groups in the early phase of the 2009 Influenza A H1N1 epidemic in both London and Birmingham [1, 2]. Similar disparities were reported in multiple settings globally [3–5]. Although there are factors unrelated to transmission which could provide explanation for measured differences in influenza related health outcomes[6–9], the analysis in chapter 3 details patterns in cases reported by ethnic and social groups which is consistent with higher rates of transmission within the South Asian population and in more deprived communities.

As discussed in chapter 2, risk of infection can be influenced by individual level factors, such as contact rate and susceptibility to infection [10]. Community-level factors such as the structure of the network of contacts surrounding an individual can also play an important role in how an infection spreads through a population. Clustering (the propensity for two contacts of an individual to also be contacts of each other) and modularity (the existence of assortative communities, where contact is more likely within the community than outside the community) in a network can have important implications for the rate at which an infectious outbreak progresses, its final size and the effectiveness of control interventions [11, 12].

In addition to impacting relative incidence over the course of an uninterrupted outbreak, community structure might also affect the ability of public health authorities to intervene and contain an outbreak in its initial phase[13]. Moreover, localised clustering of particular social or ethnic groups may introduce a perception of temporally changing relative risks as the infection spreads through communities with different ethnic or

socioeconomic composition. However, the same local clustering may also introduce inequalities if the local structure within one part of the network provides better conditions for sustained transmission than another.

Community structure within social contact networks can arise from geographical and social factors such as the concentration of contacts within towns and cities[14], with reduced contact between them[15], cultural and social divisions, where relationships may be more likely to form within social groups than between them, and through infrastructure, such as work places, transport hubs and educational institutions[16]. Large scale clustering has been modelled by making use of transport data between residential and metropolitan areas or between nations. However, identifying clustering within a single urban setting is challenging due to a lack of data on networks of contacts in particular locations or populations.

There is substantial evidence that school aged-children contribute disproportionally to influenza transmission and dynamics [17, 18] a property consistent with analysis in chapter 3. Estimates from social contact data in chapter 4 indicate that over 85% of contact exposure time between children occurs at school or at home, supporting previous findings that school and home contacts play a disproportionately large role in transmission[19]. Schools therefore provide a useful setting for studying contact structure and have been the subject of much analysis [20–25]. The body of work exploring within-school contact networks provides important insight into how children contact each other in that specific context. To complement this body of research, in chapter 5 I set out a method for using routinely collected government school data to provide insight into where children from each school live and how the schools may be linked through households

and therefore how children contact each other between schools. In the current chapter, I use the method I presented to perform simulation studies with the aim of understanding how the network structure of the school-aged population may create heterogeneity in transmission and response during an influenza outbreak in London.

London has a population of 8.7 million (2015 estimate) including 1.4 million school-aged children (16% of the population), living across 33 local authorities (Figure 6.1). The city's population has substantial socioeconomic diversity, containing both the most affluent and deprived areas of the UK. The population is also ethnically diverse with large numbers of multiple ethnic minorities resident in the Greater London area. There is a high level of geographic clustering by socio-economic status and ethnic group (Figure 6.1, Figure 1 in Appendix C). There is also substantial geographic variation in household size, which is correlated with high concentrations of certain ethnic groups, particularly ethnically South Asian populations (Indian, Pakistani and Bangladeshi). Although it is clear that these populations are clustered geographically, the strength of preferential mixing within particular ethnic or social groups is not well quantified.

In this analysis I approximated the natural community structure that arises within the school-aged population of London by building a network of contact between schools. I evaluated the local structure around each school and the implications for outbreak response. Finally, I simulated outbreaks across the network, translated cases from schools to small geographical areas to estimate incidence by socio-economic and ethnic-group from areal census data. Using the relative incidence, I assessed the implications for inequalities in risk of infection.

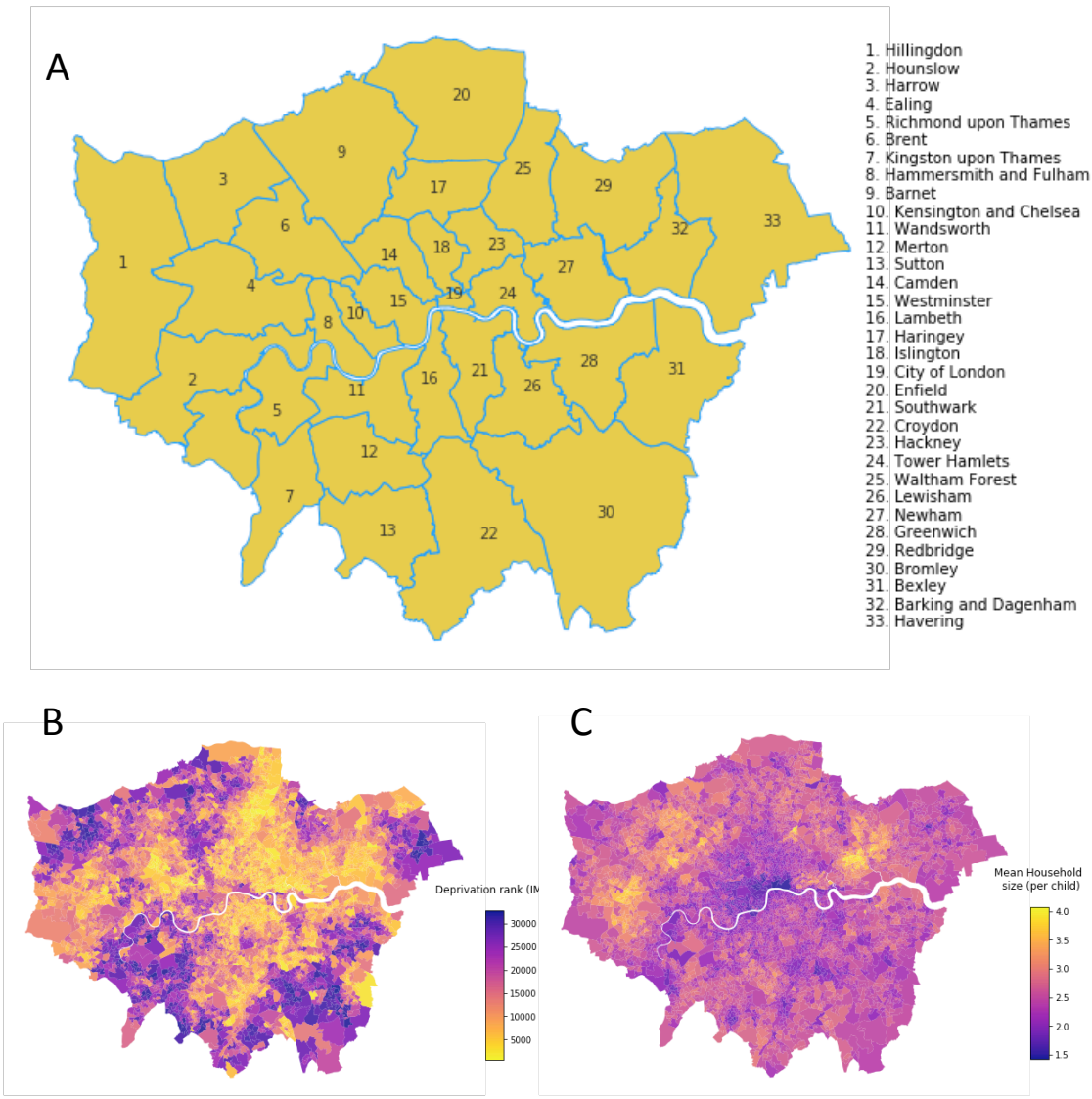


Figure 6.1 Geography, Socio-economic variation and household size in London.

A) Local authorities in London referred to as Boroughs (Names of boroughs given in key). B) Choropleth of National Lower Super Output Area (LSOA) level Index of Multiple Deprivation rank, a rank of 1 (yellow) corresponds to the most deprived the highest rank value (purple) corresponds to the most affluent. C) Choropleth of mean household size (Per child).

6.2 Methods

Summary

To evaluate how local differences in contact network structure may introduce heterogeneity in transmission, I constructed a network of schools in London using the method detailed in Chapter 5, where the edges between schools were weighted by the number of pairs of children that reside in the same household.

Explicit data for pupils who share a household was not available for London, instead I estimated the number of pairs of children that live in the same household across each pair of schools, using the rate at which children transfer from each primary school to each secondary school. I assumed that households send their children to one primary school and one secondary school. I analysed this network to evaluate heterogeneity in risk, both of experiencing an outbreak of influenza and of initiating uncontrolled outbreaks.

School data

Aggregates from the national school census were accessed via the Greater London Authority, London Data Store – Schools' Atlas Project[26]. One dataset detailed rate of transfer of children from 1843 primary schools to 559 secondary schools for the years 2014, 2015 and 2016. A second data set detailed the residence of pupils of each by Lower Super Output Area (LSOA) as the number of children in each school who live in each LSOA.

Ethnicity, socio-economic and household data

I accessed UK government census data via the UK online data service[27]. UK census is carried out every 10 years. I used Data from the 2011 census, to reflect the most recent

measurement of the population relative to the school data used. I used data aggregated at the 2011 Lower Super Output Area definition, which matches the school data aggregates. I used data detailing:

Household size: Number of households in each LSOA with 1, 2, 3, 4, 5, 6, 7, and 8 or more members.

Ethnic group: The number of people between the age of 4 and 19 who identify as each of the 19 census-defined ethnic groups.

Deprivation: National Index of Multiple Deprivation (IMD) rank. The rank of the LSOA out of all 34,753 LSOAs in England and Wales based on the IMD, a deprivation measure, which captures multiple facets of deprivation including wealth, income, living conditions, quality of life and health outcomes.

Construction of between school networks

To estimate for contact between schools, I constructed a series of networks from national school data, which I then used to analyse transmission properties of the network (Figure 6.2).

Contact network

A detailed description of this method can be found in Chapter 3, however I provide a brief description here for continuity.

The majority of schools in the UK provide either primary education for ages 4 – 11 years, or secondary education for ages 11 – 18 years. From government-collected data, I used the rate of transfer of pupils between primary and secondary schools to infer the number of pairs of contacts between schools who reside in the same household as each other. For each secondary school, *i*, the number of children transferring from each primary school, *j*, in a given year is denoted N_{obs}^{ij}.

For a household size distribution W $\{w\}$ and an average age gap between siblings of a_{gap}, the estimated number of contact pairs between schools is given by:

$$C_{ij} = \sum_{n} C_{ij}^{n} = N_{obs}^{ij} \cdot a_{gap} \cdot \sum_{n} \left[\frac{w_n (n-1)}{n} \cdot \rho_n \right]$$

Where, ρ_n is the average number of contact pairs in a household with n school-aged children:

$$\rho_n = \frac{1}{n-1} \sum_{n_s=1}^{n-1} n_s (n - n_s)$$

A detailed description of this approximation can be found in Chapter 3.

For this analysis I assumed an average age-gap between siblings of 3 years, which is the median gap between 1st and 2nd children of the same mother in the UK according to data from the United Kingdom Office for National Statistics (ONS) [27].

Approximation of the distribution of the number of school aged children per household

Census records did not provide explicit data on the number of children in the household. Instead I used household size figures to estimate this value. From this, I calculated the distribution of number of children per household unit per child.

Importantly, the estimation of contact rates between schools required me to estimate the number of children in a household at the point where children are transitioning from primary to secondary school. Under the assumption that most siblings are relatively close in age, this can be assumed to be the total number of siblings in a family. This is not directly observed in household size at a one time, as some households will be observed with fewer children than their maximum if either some children are yet to be born or some children have left home.

Firstly, to estimate the observed number of children per household from household size data, I assumed there were 2 adults per house with the rest children i.e. a household with 5 members consists of 2 adults and 3 children.

I then approximated the true distribution of number of children per household as follows. I assumed that children are generally born into and leave households one at a time (neglecting multiple birth). Also, assuming the age gap between siblings is constant, a_{gap} and that children stay in the home with parents (and other siblings) until they are A_{leave} years of age, each observation fails to identify $H_{fail,n}$ actual households with n associated children. Under these assumptions $a_{gap}(n-1)$ there are fewer than the maximum number of children in the house because some have not yet been born, and for

$a_{gap}(n\ -\ 1)$ there are fewer than the maximum number of children in the house because some have left home.

$$H_{fail,n} = H_{n,obs}\frac{2a_{gap}(n\ -\ 1)}{A_{leave}\ +\ a_{gap}(n\ -\ 1)}$$

Similarly, the observation process falsely identifies $H_{over,n}$ households with a higher total number of children (> n) as having n children before some are born and after some leave the family home. Concretely, household with more children would be observed as having n children for $2a_{gap}$ years.

Since household size is censored in the census data at 8 members, I assumed a maximum number of children to be $n_{max} = 6$ for simplicity. i.e. all households of 8 (2 adults and 6 children) or more members are taken as households of 8 members.

$$H_{over,n} = \sum_{\eta=\ n+1}^{n_{max}} \frac{a_{gap}H_{obs,\eta}}{A_{leave} + a_{gap}\ (\eta\ -\ 1)}$$

Hence, I estimated the number of households with n children as:

$$H_n = H_{obs,n} + H_{fail,n} - H_{over,n}$$

$$= H_{obs,n}\left(1 + \frac{2a_{gap}(n\ -\ 1)}{A_{leave} + a_{gap}(n\ -\ 1)}\right) - \sum_{\eta=\ n+1}^{n_{max}} \frac{2a_{gap}H_{obs,\eta}}{A_{leave} + a_{gap}\ (\eta\ -\ 1)}$$

For this analysis I have used an age gap of 3 years between siblings, which is the median interval between births to the same mother in the UK. I assumed children leave home at the age (A_{leave}) of 18 years old.

Transmission probability network

I used the contact network to estimate the probability of transmission between each school experiencing an outbreak and each of its adjacent schools assuming density dependent transmission within households.

I equated the probability of transmission between schools as the complement of the probability that transmission does not occur between any pair of pupils who share household, assuming each pair has the same probability of transmission between them:

$$P_{trans,ij} = 1 - \left(1 - P_j^I P_i^S q P_i^{OB}\right)^{C_{ij}}$$

The contact in an infected school, j, had probability of P_j^I of being infected. I assumed this to be the proportion of the school infected during an outbreak, approximated by solving the final size equation:

$$P_j^I = R_j(\infty) = 1 - e^{-R_0 R_j(\infty)}$$

Where, R_0 is the basic reproduction number for school-based transmission and $R_j(\infty)$ is the final size of an outbreak (proportion recovered at time $t = \infty$) in school j.

Each contact in adjacent school i had probability P_i^S of being susceptible, assumed to be 1 in this analysis to reflect an unvaccinated population. Transmission between children within the household occurs with probability q. The probability of an outbreak in school i in the event of a single child becoming infected was P_i^{OB}.

$$P_i^{OB} = \left(1 - \frac{1}{R_0}\right)$$

For the purposes of this analysis, to simulate an outbreak similar to Influenza A H1N1 09, I used R_0 values between 1.1 and 2 based on previous analysis of R_0 during 2009/2010 [28–30]. I chose a probability of transmission between siblings, q, of 0.15, based on the results of a detailed analysis of transmission between siblings in Japan [31].

Under this framework, an outbreak was seeded in a particular school which I refer to as the *index school*, the outbreak can spread between schools, where infected schools *seed* outbreaks in adjacent schools with probability $P_{trans,ij}$.

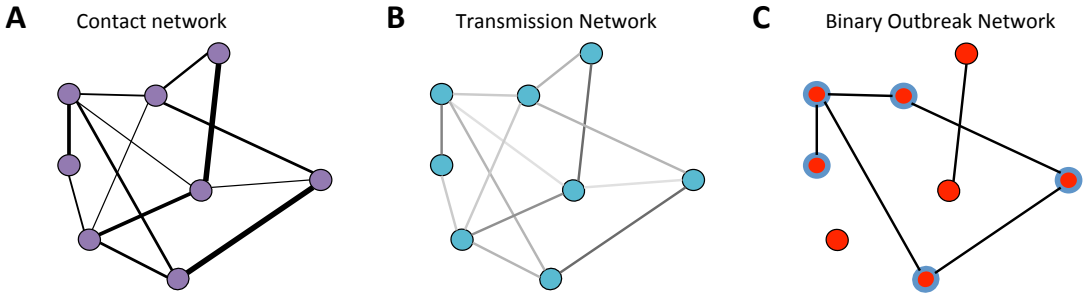


Figure 6.2 Contact networks, transmission networks and outbreak networks

A) A schematic of a contact network, the width of the edges shows the relative probability of transmission between schools B) A schematic of a transmission probability network calculated from the contact network, the shading of the edges shows the relative probability of transmission between schools. C) A schematic of a realisation of a binary outbreak network (sampled from A), where edges are weighted 1 with probability given by the equivalent edge in the transmission network, or 0 otherwise. Blue highlighted nodes show those in the largest connected component. In each network nodes show the location of schools.

Binary outbreak network

To estimate the risk of an outbreak in each school I simulated multiple realisations of outbreaks on the network according to the calculated probabilities between schools. Since the proportion susceptible in each school was assumed to be 1, the probability of transmission in each direction is equal; hence the network in each realization can be approximated as a non-directed *binary outbreak network* (Figure 4.2), where edges were weighted 1 or 0. An edge weight of 1 indicates that transmission would occur between these schools in the event of an outbreak in one of them; an edge weight of 0 indicates no transmission occurs between schools. Edges were weighted 1 with a probability of $P_{trans,i,j}$. Creation of such networks was computationally faster than using an iterative "diffusion" method progressing across the network from a particular school because although every edge in the network must be evaluated, it allows every school to be analysed with little additional computational effort in a single network realisation.

Describing the network

I calculated the degree (the number of connected schools and weighted degree) number of contact pairs, of each school in the contact network. I evaluated how these varied geographically and by socioeconomic status. To assess the relationship to household size I used Spearman's rank to quantify the correlation between estimated mean number of children per household represented in the school and the degree and weighted degree.

Transmission characteristics of individual schools

To assess the relative ability for public health authorities to contain an outbreak originating in each school across the network, I used the transmission probability network to calculate two school level transmission metrics:

1. Mean number of outbreaks seeded in adjacent schools; if a large number of schools are infected by the index school it is likely that contact tracing will be more challenging.
2. Mean proportion of students infected before seeding an outbreak in an adjacent school; if only a small proportion of the school is infected it is less likely that authorities would identify the outbreak before a second school is seeded.

I used the transmission probability network of schools (Figure 6.2 B) to calculate each of these metrics for each school.

Mean number of outbreaks seeded in adjacent schools

To establish the expected number of adjacent schools infected in the event of an outbreak in each individual school, I calculated the sum of the probabilities of transmission to all adjacent schools. This is simply the weighted degree of the school in the probabilistic transmission network.

Mean proportion of students infected before seeding an outbreak in an adjacent school

To estimate the expected proportion of the school infected by the time the first outbreak was seeded in an adjacent school, I calculated the proportion of children infected which would lead to a 50% chance of infecting another school (i.e. $P_{trans,ij}(t) = \frac{1}{2}$).

$$P_{trans,ij}(t) = 1 - \left(1 - p_{inf} q P_i^{OB}\right)^{c_{ij}} = \frac{1}{2}$$

Hence,

$$p_{inf} = \frac{1}{q P_i^{OB}} \left(1 - \frac{1}{\sqrt[c_{ij}]{2}}\right)$$

Where, p_{inf} is the proportion of the school infected by the time to seeding is expected.

Risk of infection by school

Connected components

To establish each school's risk of infection over the course of an uninterrupted outbreak, I identified connected components of 1000 unique realisations of a *binary outbreak network*. A connected component is a collection of schools that are all connected by at least one network path (chain of edges weighted 1). For a given connected component of the *binary outbreak network*, every school will infect every other school in the component when seeding an outbreak. Likewise, each school would be infected if an outbreak is seeded by any other school in the component. Complementarily, a school cannot be infected by, or infect any school outside their connected component. Each school is either part of only one connected component, or is unconnected to any node.

I quantified risk of infection to children in each particular school as the proportion of children in the network who attend the schools in the connected component to which their own school belongs, i.e. the proportion of children who could initiate an outbreak across the school network that leads to their infection. I repeated the analysis for values of R_0 ranging from 1.1 to 2, which is considered consistent with estimated R_0 for an influenza outbreak in the UK[28–30].

Establishing inequalities between social groups

To assess the geographic, socio-economic and ethnic variation in risk, I estimated the risk to children living in each specific LSOA by total weighted risk in school children living in the LSOA in attendance. i.e. I weighted each schools risk by the proportion children living in the LSOA. Using national census data (2011) on LSOA level deprivation status and the number of school-aged-children from each ethnic group in each LSOA; I calculated mean risk for each quintile of the Index of Material Deprivation and ethnic group. For each social group (ethnic and socio-economic) I calculated the relative risk as the ratio of the individual groups risk and the mean risk for the whole network.

Changing relative risks over the course of an outbreak

Due to clusters with a high concentration of particular ethnic and socio-economic groups in certain parts of the network, the distribution of cases by ethnic group and socio-economic status may vary over the course of the outbreak depending on the school where the outbreak initiates. To evaluate how inequality in incidence varies depending on the index school, I estimated the relative risk of infection in each ethnic group and deprivation quintile at the first 15 generations of an outbreak, where a generation is defined such that the schools infected by the index school form the first generation, schools infected by the first generation form the second generation etc. Relative risks at each generation were calculated using the cumulative incidence in schools infected up to and including the generation considered. I evaluated the changing relative risk over the course of outbreaks initiated in every secondary school in the network.

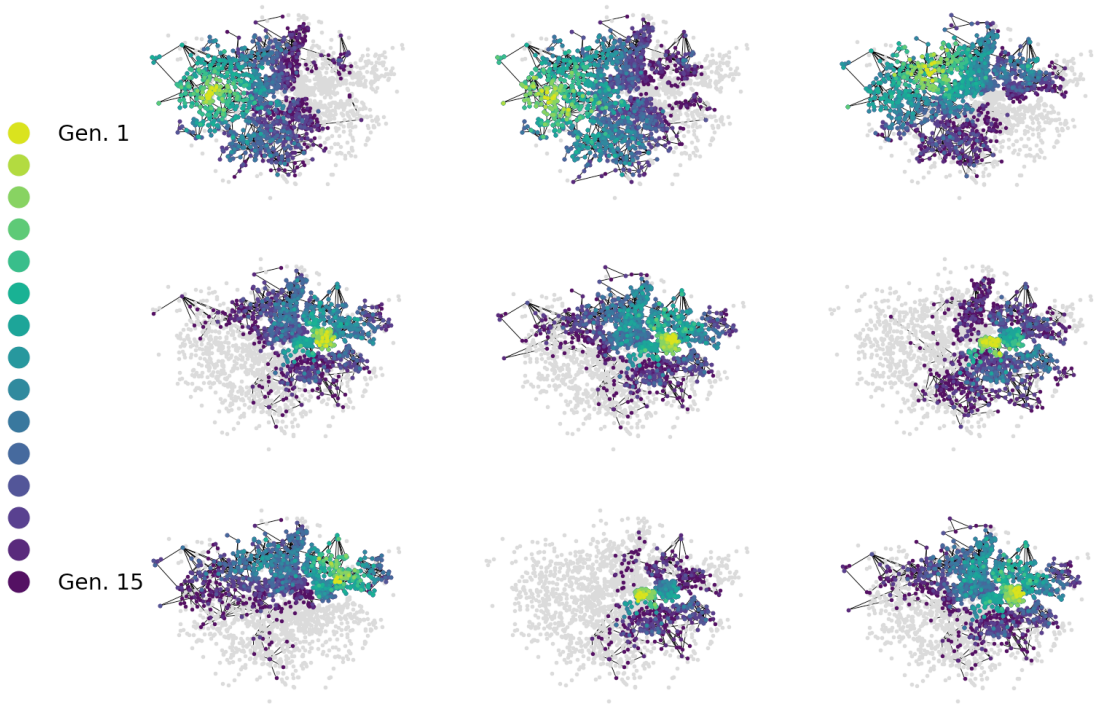


Figure 6.3 Graphs of the first 15 generations of outbreaks in the 9 schools with the highest weighted degree, for a given sampled binary outbreak network with R_0 of 1.5.

coloured nodes indicate the location of schools in successive generations of the outbreak (yellow to purple). Grey nodes indicate the location of schools not involved in the first 15 generations of the outbreak.

To establish each progressive generation of infection I extracted ego-networks (networks centred around a particular 'index' school) from unique realisations of a *binary outbreak network*. I evaluated the ego-networks around every secondary school in the network with radius 1 to 15, where the radius is the minimum number of edges that link the index school and the most distant schools in the ego-network.

To estimate the ethnic and socio-economic distribution of infection after each generation of an outbreak, I calculated the proportion of children in each ethnic group and deprivation quintile at each network radius. From this I estimated the relative risk of infection in each ethnic and deprivation quintile as the ratio of the proportion of the

infected schools that belong to each group to the proportion of the total school population of London belonging to the same group.

Sensitivity of the network to variation in R between schools

The transmission network presented in this analysis relies upon two key epidemiological parameters. Firstly, R_0 which describes the strength of transmission within schools and secondly q which defines the transmission probability between members of the same household. In this section I examine the implications of the way these parameters are implemented in the model.

Firstly, the approach I used to simulate outbreaks required R_0 to be equal across all schools buts this is unlikely to be the case. R_0 is not a fixed value even within a particular population and is therefore unlikely to be fixed even within one school over time. To evaluate the impact of variation in R_0 between schools, I applied four regimes of variation, two with un-correlated variation, where R_0 is A) normally and B) log-normally distributed across the network, and two where R_0 is dependent on the 'phase' of the school (primary or secondary). All regimes were evaluated with a mean R_0 of 1.5 and 1.3 with a coefficient of variation of 0.133. deviation in R_0

Regime	Mean R_0 = 1.5	Mean R_0 = 1.3
Normal	R_0 ~ NORM(1.5, 0.2)	R_0 ~ NORM(1.3, 0.178)
Lognormal	R_0 ~ LOGNORM(1.5, 0.2)	R_0 ~ LOGNORM(1.3, 0.178)
Higher in Primary schools	R_{Pri} = 1.55, R_{sec} = 1.35	R_{Pri} = 1.35, R_{sec} = 1.15
Higher in Secondary schools	R_{Pri} = 1.45, R_{sec} = 1.25	R_{Pri} = 1.65, R_{sec} = 1.45

Table 6.1 Sensitivity analysis regimes for variation in R between schools

I evaluated the impact on the degree distribution of the transmission probability network and the expected component size of a few select schools. In the framework I present, allowing R_0 to vary between schools, results in different transmission probabilities in different directions along the same edge of the network. The transmission network can therefore only be fully expressed as directed graphs. As a result, a connected component of the, now directed, binary network no longer indicates the outbreak size expected from each school in that component. Instead, I evaluated the outbreak size by taking the full chains of 'successors' (more detailed discussion of successors in the methods section of chapter 8 of this thesis) on the directed graph from a given school for each of the regimes, over 100 realisations of the binary outbreak network (equivalent to 201,100 outbreak simulations). I used the components to evaluate the relative risk of infection in each deprivation quintile and ethnic group as described in the main analysis.

Secondly, In the main analysis I chose a probability of 0.15 between members of the same household. To evaluate the impact of changing this parameter value, I have calculated the overall relative risk of infection by deprivation quintile and ethnic group with half (q=0.08) and double (q=0.3) the probability of transmission between members of the same household.

6.3 **Results**

School contact network

The contact network between schools contained 2,027 schools, of which 1,954 of them were in the largest component of the network. The remaining schools were in effect disconnected from the network and in components of three (n. 1), two (n. 23) and one (n.

24) schools. The mean degree of the contact network, the average number of adjacent schools with at least one contact pair, is 4.5 schools. The mean weighted degree of the contact network, the average number of contact pairs with other schools, is 205.3.

Degree distributions vary between schools located in different boroughs (Appendix C Figure 2). In particular the Boroughs of Newham, Redbridge, Waltham Forest, Tower Hamlets and Brent had broader distributions, with generally higher number of contact pairs between schools. The borough with the highest mean number of contact pairs per school was Newham, 442 The borough with the lowest mean number of contact pairs per school was City of London, 49. The school with the lowest number of contact pairs was a primary school in the borough of Merton with 8.72 unique contact pairs. The school with the largest number of contact pairs was a secondary school in the London Borough of Newham and had 1900 unique contact pairs connecting it to adjacent schools.

The weighted degree of each school was moderately correlated with mean household size (Figure 6.4) with a Spearman's rank (SR) of 0.34 ($p << 0.01$). The association between weighted degree and household size was stronger amongst primary schools (SR = 0.66 ($p<<0.01$)) than secondary schools (SR=0.32 ($p<<0.01$)). The degree, or number of unique connected schools, was also weakly correlated with household size with a Spearman's rank of 0.12 ($p << 0.01$), but more strongly for primary schools amongst which Spearman's rank increases to 0.5 ($p << 0.01$), indicating that regions with large households have higher connectedness within the school network.

Transmission characteristics of individual schools

To quantify the rate of transmission at a school level, I calculated three metrics for each school based on their weighted degree, the number of unique contact pairs with adjacent schools in the network.

Expected number of adjacent schools infected

The expected number of adjacent schools infected was strongly correlated with the weighted degree of each school at low values of R_0 (Spearman's Rank 0.99 ($p<<0.01$) at R_0 = 1.1), reducing in correlation as R_0 increased (Spearman's Rank 0.87 ($p<<0.01$) at R_0 = 2). Conversely the expected number of adjacent schools infected was most strongly correlated with the degree at higher values of R_0 (Spearman's Rank 0.98 ($p<<0.01$) at R_0 = 2), with weaker correlation at low values (Spearman's Rank 0.82 ($p<<0.01$) at R_0 = 1.1) (Appendix C Figure 3). This suggests that the number of adjacent schools has more influence on the number of seeded outbreaks at higher R_0 and the number of between-children contacts is more important at low values of R_0.

The mean expected number of adjacent schools infected increased with R_0 from 0.35 with R_0 of 1.1 to 3.55 with R_0 of 2. The average was higher in secondary schools than primary schools, with a secondary school mean of 6.76 with R_0 of 2 compared to a primary school mean of 2.42.

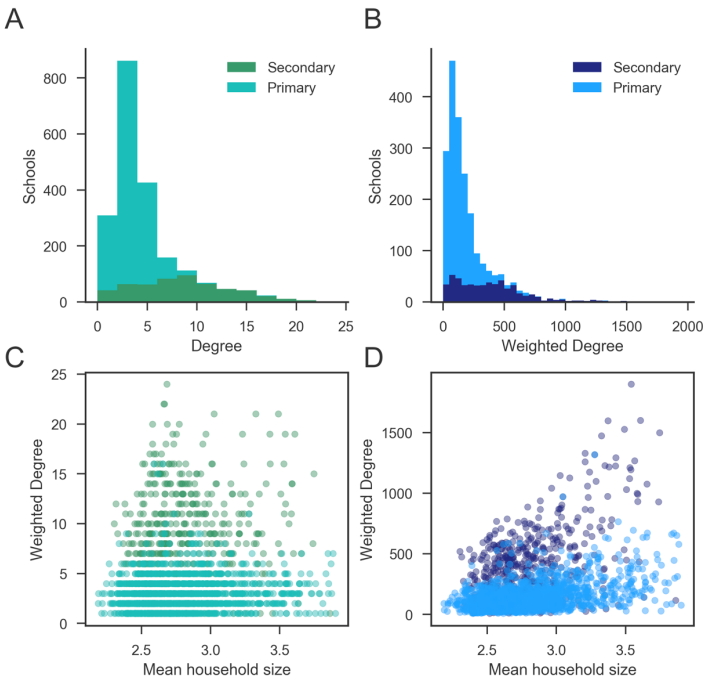


Figure 6.4 The degree (A) and weighted degree (B) distributions of the between school contact network constructed from National School Census data.

The degree of a school is the number of schools adjacent to it in the network irrespective of the number of contact pairs. The weighted degree is the number of contact pairs between a school and all its adjacent schools. Relationship between degree (C) and weighted degree (D) and the mean household size of children each school.

As the expected number of adjacent schools infected was strongly correlated with the weighted degree of the between school contact network, similar geographic variation was present to the mean number of unique contacts (Appendix C Table 1). East London boroughs of Newham, Redbridge and Tower hamlets were consistently highest. The school with the highest expected number of adjacent schools infected, was a secondary school in Tower Hamlets, expecting 39.6 outbreaks in adjacent schools on average when R_0 was 2.

In general, South Asian children were most likely to attend schools that infect adjacent schools than average, with mean expected adjacent schools infected 16%, 19% and 36%

higher for children of Indian, Pakistani and Bangladeshi ethnicity respectively, with a within school R_0 of 1.1. This relative difference reduced with increasing R_0, to 4%, 6% and 21% higher with a within school R_0 of 2.0 (Figure 6.6).

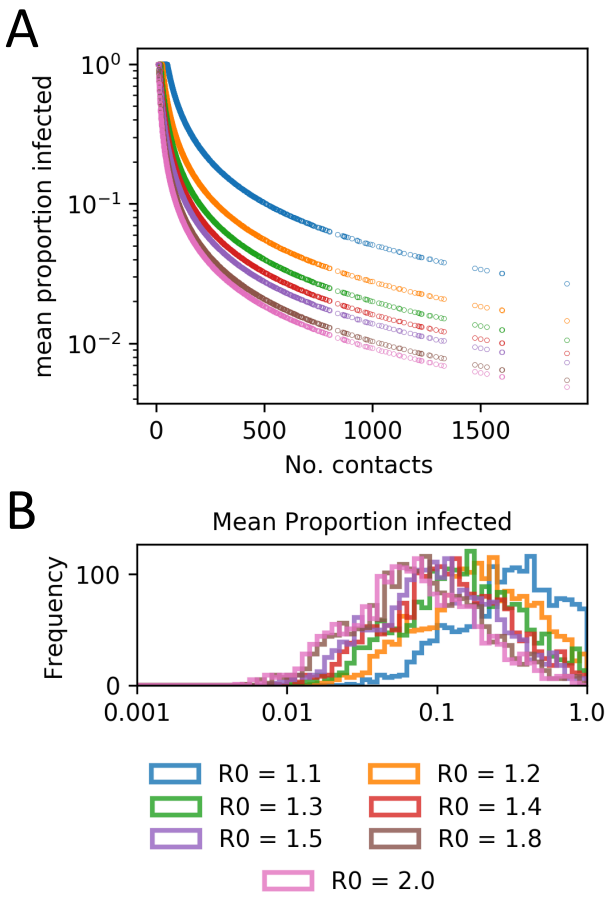


Figure 6.5 A) Proportion of school infected before seeding an outbreak in an adjacent school plotted against the weighted degree of the between school contact network. B) Histogram of the proportion of school infected before seeding an outbreak in an adjacent.

Expected proportion infected before onward seeding

For a particular value of R_0 and q (the probability of transmission between household members), the expected proportion infected before onward seeding is dependent only on the total number of unique contact pairs with adjacent schools. It therefore monotonically decreases with the weighted degree of the school in the school contact network (Figure 6,5).

The median expected proportion of the school population infected before the outbreak on the school seeds an outbreak in an adjacent school was 46% with an R_0 of 1.1 compared to an estimated final size of 18% infected. This proportion dropped to 9% with an R_0 equal to 2, compared to a final size of around 80%. A secondary school in Newham had the minimum value; with R_0 equal to 2 the estimated proportion infected was 0.4% (approx. 5 pupils) before starting an outbreak in another school. There were a number of schools in the network which had a probability of seeding a second outbreak of lower than 0.5, even when the whole school was infected.

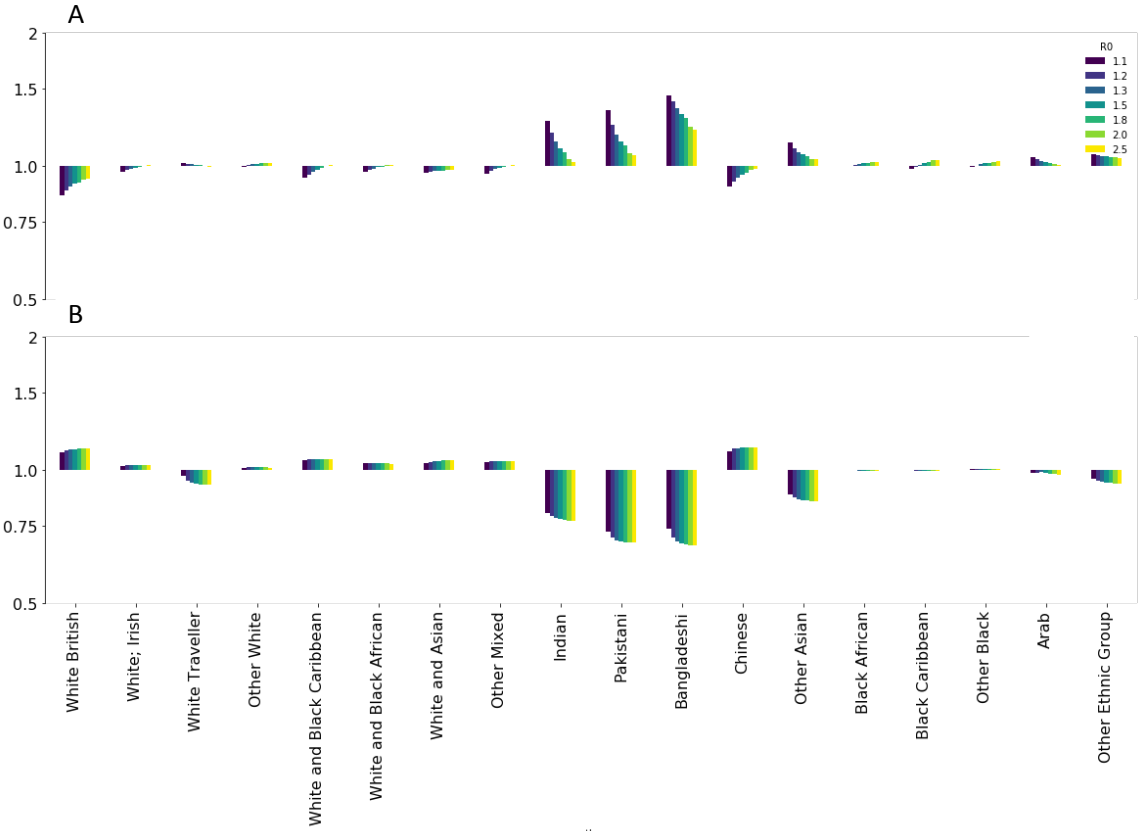


Figure 6.6 (A) Relative magnitude of expected number of adjacent schools infected, (B) and mean proportion infected before seeding a second outbreak, by ethnic group, for values of R_0 from 1.1 to 2.

On average, children of South Asian ethnicities attended schools with a proportion infected before seeding an outbreak in an adjacent school, between 29% and 35% lower than average for R_0 between 1.1 and 2 respectively (Figure 6.6). Similarly, children living in more deprived LSOAs on average attended schools which had a lower proportion of the school infected before onward seeding would be expected (Figure 6.7). However, the difference between deprivation quintiles was not as clear as between ethnic groups.

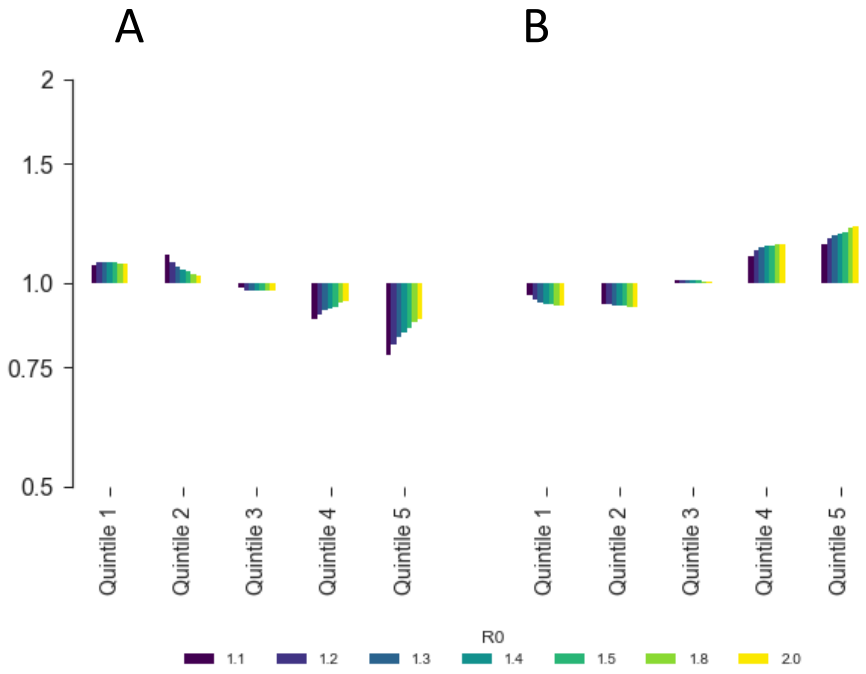


Figure 6.7 (A) Relative magnitude of mean number adjacent schools infected. (B) Mean proportion infected before seeding a second outbreak, by deprivation quintile, for values of R_0 from 1.1 to 2.

Overall risk of infection

Over all realisations of the binary outbreak network, the size of the largest component of the binary outbreak networks increased rapidly at low values of R_0 (Figure 6.8). From a mean of 19.5 schools (1% of schools) and interquartile range of 16 - 23 schools at R_0 of 1.1 to a mean of 1571.7 (78% of schools) and interquartile range of 1562 – 1604 at R_0 of 1.4. At an R_0 of 2.0 the largest component had a mean of 1871 (92% of schools) and

interquartile range of 1865 - 1877. For all R_0 the majority of schools outside of the largest component were in small components of less than 10 schools (Figure 6.8).

	Largest component			
R0	mean	lower quartile	median	upper quartile
1.1	19.5	16	19	23
1.2	192.0	144	174	224
1.3	1105.2	919	1178	1279
1.4	1571.7	1562	1585	1604
1.5	1713.6	1703	1713	1724
1.8	1840.7	1835	1840	1847
2	1871.0	1865	1871	1877

Table 6.2 The largest component of Binary Outbreak Networks calculated over 1000 realisations,

The change in the size of the largest component at low values of R_0 meant that variation in overall risk of infection was also highly dependent on the value of R_0 (Figure 6.9). At low values of R_0, between 1.1 and 1.3, there is substantial variation in risk across the network, resulting in inequalities by London borough, ethnic group and socio-economic status.

Schools attended by children living in North East London boroughs of Tower Hamlets, Newham, Redbridge and Enfield were found to be at particularly high risk. West London boroughs of Harrow, Hounslow, Ealing and Brent also have higher risk than the London mean.

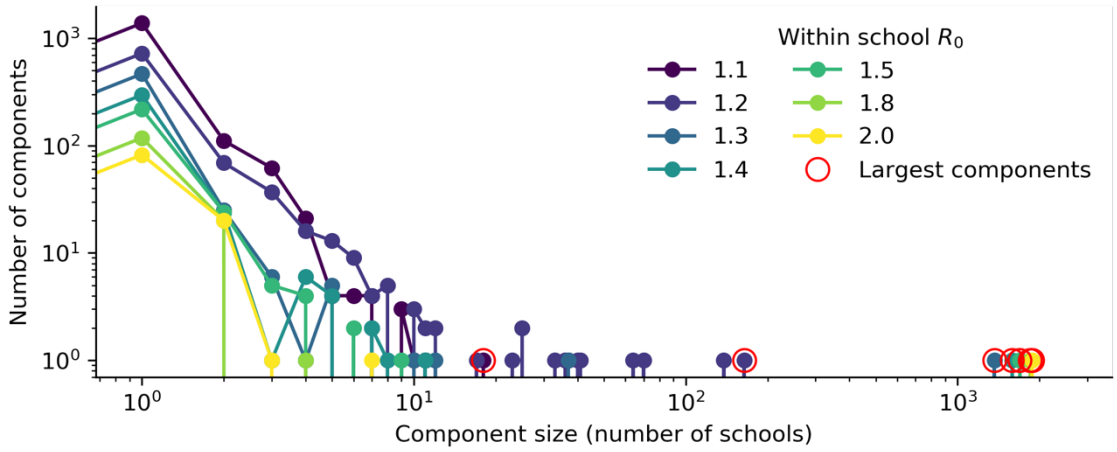


Figure 6.8 Example component size distribution of binary outbreak networks

with R_0 of 1.1 to 2.0. points show counts of components with x schools. Colour indicates R_0. Red rings show the largest component for each R_0 value.

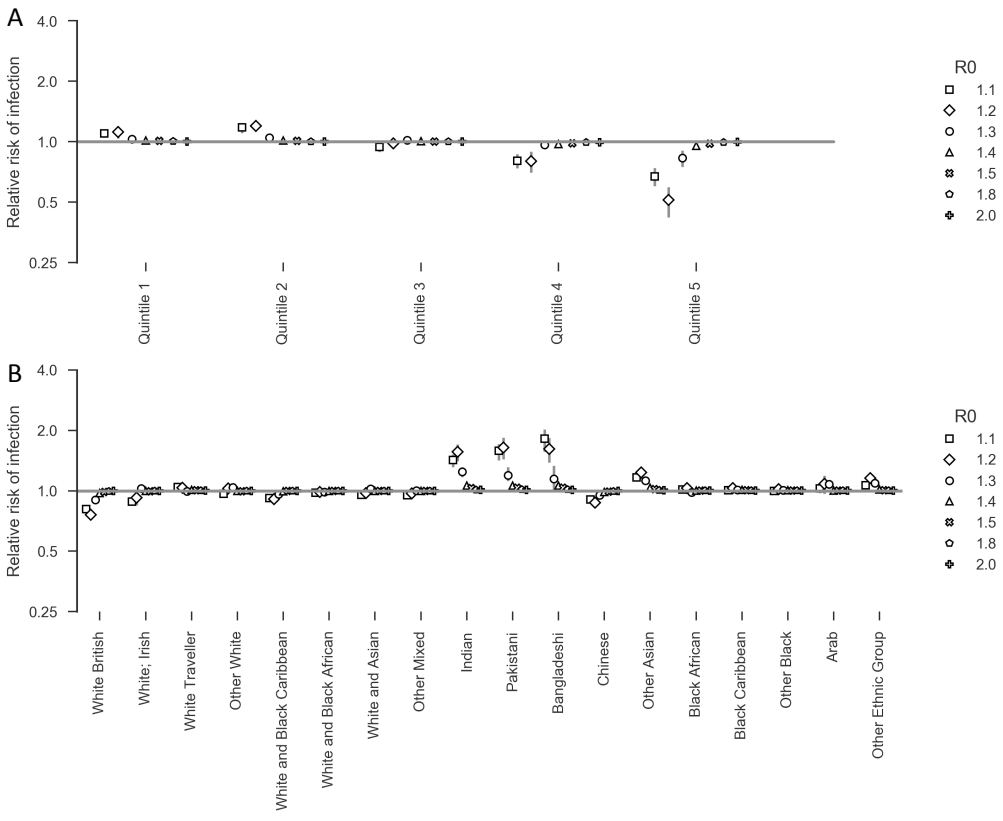


Figure 6.9 The relative risk of infection in outbreaks on the school network

Relative risk of infection by A) deprivation quintile and B) ethnic group. Markers show the overall relative risk of infection during an outbreak initiated in any school on the network at R_0 values from 1.1 to 2.0.

Similar to school level metrics above, there is notably higher than expected risk of infection in south Asian children with a peak relative risk of 1.58, 1.66 and 1.61 in Indian, Pakistani and Bangladeshi children respectively with R_0 of 1.2 (Figure 6.9). This coincides with lower than expected incidence in white British children (RR 0.75). There is also notably lower risk (RR 0.48 at R_0 = 1.2) in the most affluent quintile at the same, lower values of R_0. The variation in overall risk diminishes at higher values of R_0, with similar distribution in risk of infection in schools across London, regardless of geographic location, deprivation status or ethnic composition.

Changing risk of infection in the early phase of an outbreak

To assess how relative risk of infection varies over the first 15 generations of an outbreak with each secondary school as the index school, I extracted ego-networks from a binary outbreak network with each secondary school as the index case. I calculated the relative risk of infection by ethnic group and deprivation quintile after each generation of an outbreak for ego-networks of radius 1 to 15. Early in the outbreaks, all ethnic groups had instances of either disproportionately high or low risk, depending on the index school. South Asian populations (Indian, Pakistani and Bangladeshi) showed the greatest deviation from equitable risk of infection, with much reduced and increased risk at the beginning of the outbreak (generation 1). Bangladeshi children had maximum relative risk of around 16 and minimum relative risk of 1/16 (Figure 6.10).

In most ethnic groups, the risk of infection became proportionate within the first 15 generations for almost all outbreaks. However, in South Asian children, a large number

of outbreaks were simulated where the risk remained either disproportionately high or disproportionately low.

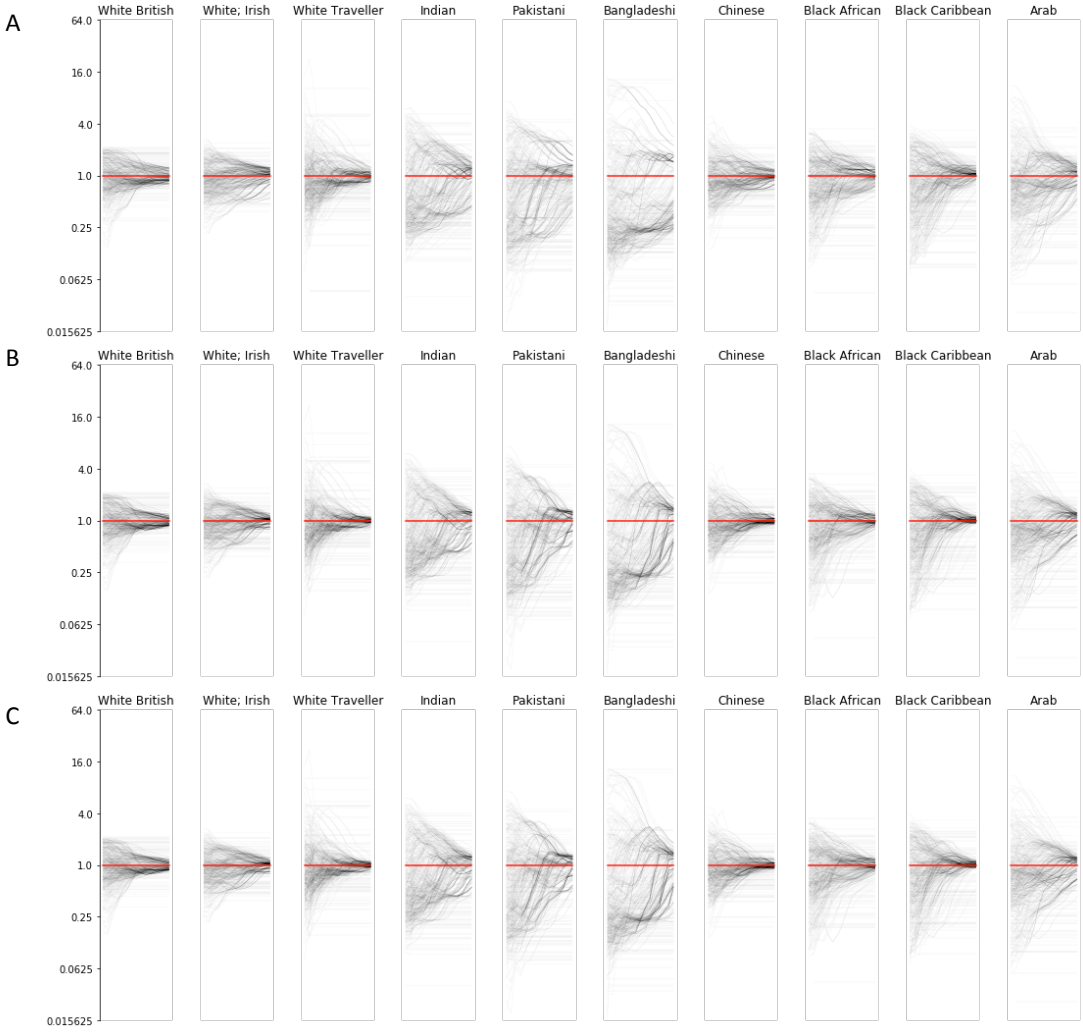


Figure 6.10 Relative risk by Ethnic Group in the first 15 generations (of schools) of outbreaks seeded in each secondary school in the Network

with R_0 of 1.5 (A), 1.8 (B) and 2.0 (C). Grey lines show the relative risk in each ethnic group at progressive generations of outbreaks originating in each school. The red lines show a relative risk of 1, which indicates proportionate risk for each ethnic group

Similarly, the risk of high deviation from equal distribution of infection between deprivation quintiles was greatest early in the outbreak (Figure 6.11). The greatest deviation was observed in the most affluent and most deprived quintiles, with intermediate quintiles showing lower potential inequalities.

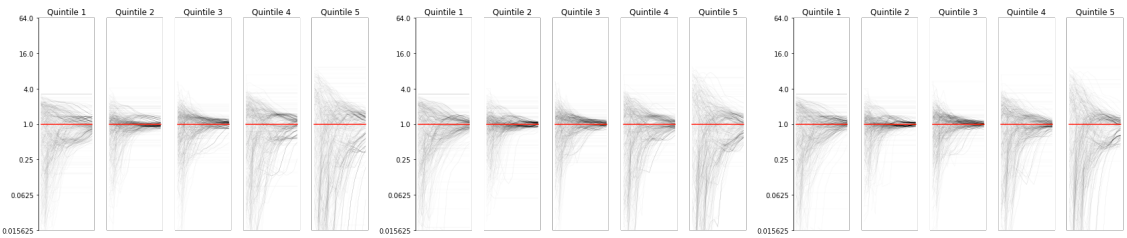


Figure 6.11 Relative risk by deprivation quintile in the first 15 generations (of schools) of outbreaks seeded in each secondary school in the Network

with R_0 of 1.5 (A), 1.8 (B) and 2.0 (C). Grey lines show the relative risk in each deprivation quintile at progressive generations of outbreaks originating in each school. The red lines show a relative risk of 1, which indicates proportionate risk for each deprivation quintile.

Sensitivity of the network to variation in R between schools

At a high level, introducing variability in R_0 had minimal impact on the network. The degree distributions of the transmission probability network remained comparable for all regimes. Unstructured variation in R_0 between schools resulted in a slight reduction in mean weighted degree, for example for a mean R_0 of 1.3, from 1.85 (1.76, 1.95; 95% CI) with no variation to 1.57 (1.48, 1.67; 95% CI) and 1.56 (1.47, 1.65; 95% CI) for normally and log-normally distributed R_0 respectively. Increasing R_0 in primary schools and reducing it in secondary schools reduced the mean degree to 1.36 (1.29, 1.44; 95% CI), whereas reducing R_0 in primary schools and increasing in secondary schools increased the mean degree to 1.98 (1.88, 2.08; 95% CI).

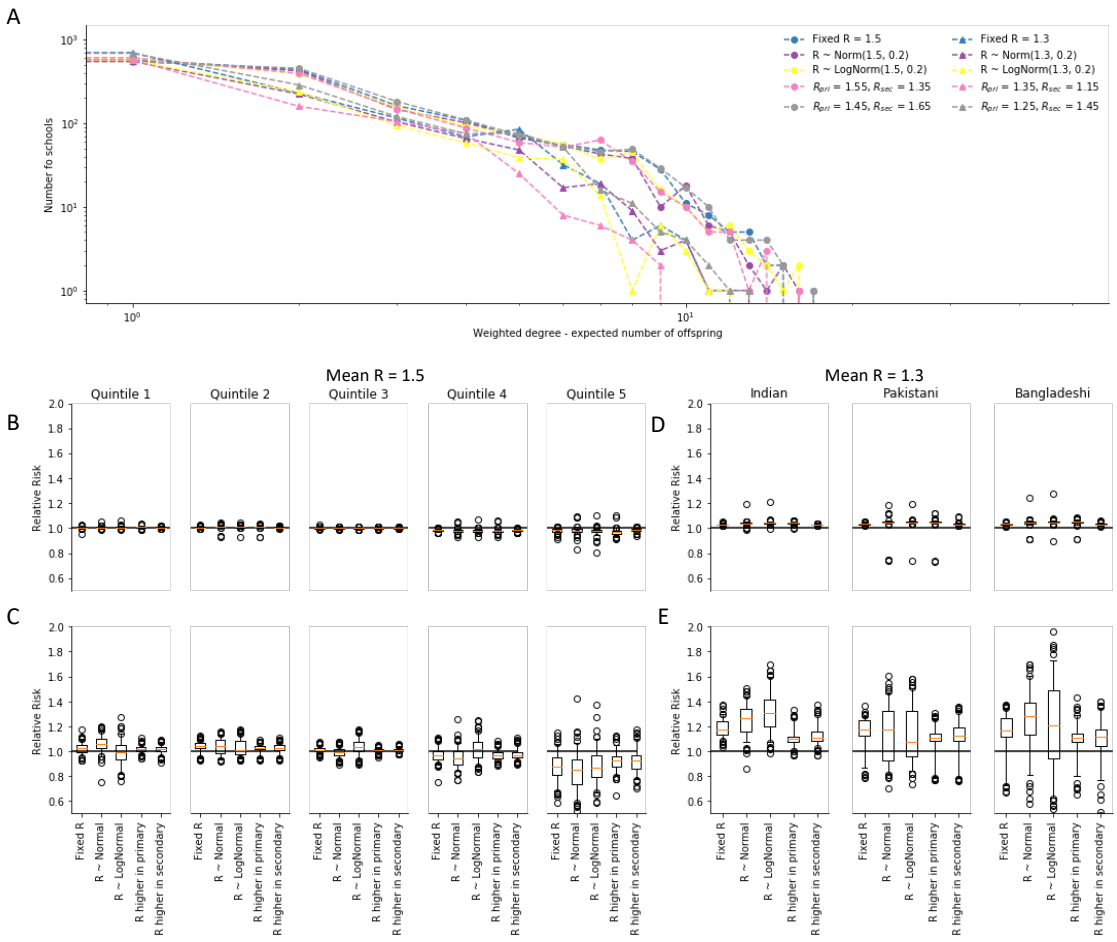


Figure 6.12 Sensitivity analyses – Variation in R_0 between schools

A) Weighted degree distributions of the transmission probability network for each regime of variation in R_0. Relative risk of infection by deprivation quintile for mean R_0 of 1.5 and standard deviation of 0.2 (B) and mean R_0 of 1.3 and standard deviation of 0.178 (C). Relative risk of infection by ethnic group (South Asian ethnicities) for mean R_0 of 1.5 and standard deviation of 0.2 (D) and mean R_0 of 1.3 and standard deviation of 0.178 (E). Combinations of R_0 for primary and secondary schools with mean R_0 of 1.5 and 1.3 are shown in Table 6.1.

The introduction of variation in R_0 had very little impact on the inequalities calculated for networks with a mean R_0 of 1.5 (Figure 6.12 (B and D)). Changes in relative risk were greater for mean R_0 of 1.3, particularly for inequalities by ethnicity where uncertainty in relative risk increased markedly when R_0 varied randomly (both normally and log-normally distributed, however the range of values remained similar. Notably the lower

quartile of relative risk values simulated straddle equity (relative risk equal to 1.0) for Pakistani ethnicity with normal distributed R_0 and for Pakistani and Bangladeshi Ethnicities for log-normal distribution of R_0. Despite these differences the results remain qualitatively similar. Assigning different values of R_0 to primary and secondary schools reduced relative risks in all ethnicities marginally, with little change to the variation between realisations of binary outbreak networks.

Sensitivity of the network to the within-household transmission probability (q)

Changing the probability of transmission between household members had a greater impact on relative risks at low values of R_0. In essence, reducing transmission probability had the impact of shifting the results such that they were more similar to lower values of R_0 with the original values of q (Figure 6.13). Complementarily, when q was increased the results were shifted such that they were more similar to higher values of R_0. Typically for values of R_0 greater than 1.2 reducing q increased inequalities, whilst increasing q reduced them. Broadly the overall values of relative risk remained similar across the whole range of values of R_0, but the values of $R_0$to which they corresponded changed .

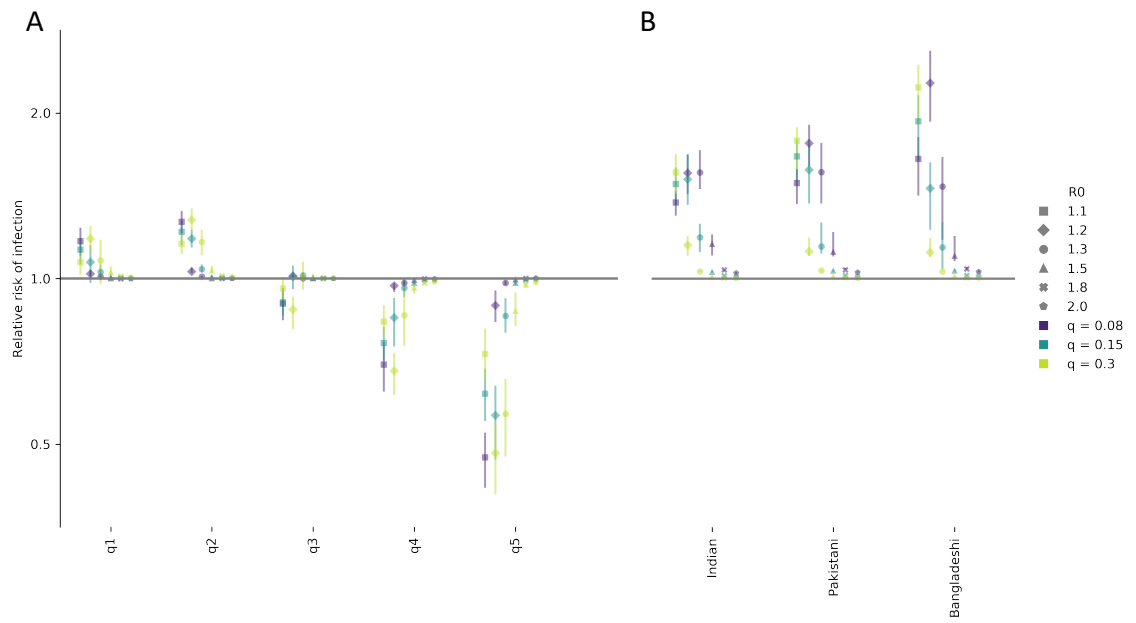


Figure 6.13 Sensitivity analyses – within-household transmission probability

The relative risk of infection by deprivation quintile (A) and ethnicity (B) over the full range of R_0 and q. Markers show the mean relative risk, bars show the interquartile range.

6.4 **Discussion**

Inequalities in health outcomes have been observed during outbreaks of influenza in multiple settings. Community structure within contact networks can have important implications for transmission due both to clustering of particular groups and to providing a means for local differences in network properties to introduce inequalities in risk of infection.

I used UK government collected school and census data to construct a network of schools linked by children who live in households with children attendant at other schools. I analysed the community structure within a contact network and its implications for variation in risk of transmission between schools across the network. In particular I used

the network to assess whether the network possesses properties that promote higher risk of infection in certain socio-economic and ethnic groups.

The school to school contact network revealed systematic differences in local network properties by geographic location, socio-economic and ethnic group, with higher number of household contacts linking schools in more deprived areas and in areas with a high proportion of South Asian children. The variation in contact between schools lead to differences in transmission related factors, which could contribute to inequalities in outbreaks.

Firstly, I assessed the properties of each school to evaluate the relative onward transmission of infection following a school level outbreak. The higher rate of contact between schools in particular groups of the population caused an increase in the expected number of adjacent schools infected per school and a lower expected proportion infected before seeding a second outbreak.

Each of these may impact the rate of transmission across school boundaries. In particular, South Asian children generally attended schools with higher rates of contact with other schools, however the relative differences in mean values were quite small (RR < 2.0). This relatively small difference makes it difficult to ascertain how much this may influence inequalities in reality. This is dependent firstly on how much the metrics will impact ability of Public Health Authorities to respond, and secondly how effective any intervention available to public health authorities may be in containing an outbreak if discovered early. Importantly, because these factors are correlated, they compound each other. Notably, there were 25 outlying schools that had between 1000 and 1850 unique

contacts with children in adjacent schools. A number of these had disproportionately high South Asian school populations. If there is significantly higher risk of an uncontrolled outbreak within these schools due to their connectedness, South Asian children would be disproportionately likely to be infected early in such an outbreak.

Secondly, I assessed the overall risk of infection in each school to establish relative risk of infection to each ethnic group and deprivation quintile. At very low values of R_0 (1.1 – 1.3), where outbreaks remain small, the risk posed to South Asian populations was higher. This is due to outbreaks in these areas generally having a higher final size, giving children in these populations a greater risk of infection overall. This disparity in overall risk was diminished at higher values of R_0 because outbreaks were expected to be much larger covering the whole geographic area of London, suggesting that in an uninterrupted outbreak on the network, pupils of all ethnicity and socioeconomic status are at a similar risk of infection at school.

However, when I evaluated disparities in incidence over the course of the first 15 generations of an outbreak, inequalities are more likely to be measured early in the outbreak. In particular, South Asian populations are highly clustered within the network. The increased variability in relative risk in South Asian children indicates that larger relative risks may be measured in this group relative to other ethnic groups in the population. The presence of this inequality is dependent on seeding, if an outbreak effects a South Asian population early in the epidemic, it is likely that large inequalities will be measured in the initial phase, whereas if no infection is introduced into that population early in the epidemic, reduced risk will be measured in this group.

It is possible that although there is no indication from the network that an uncontrolled outbreak would disproportionately impact South Asians at values of R_0 over 1.3, the combination of lower containability and the potential for large inequalities means that South Asians may be at higher risk of infection in the early stages of an influenza outbreak. In order to determine this the ability for public health authorities to respond and contain a school outbreak would need to be quantified.

A possible explanation for repeated reports of outbreaks with higher incidence in minority groups is that of confirmation bias. Instances where outbreaks have reached populations of particular ethnicity early in the outbreak may have been highlighted due to the perceived presence of inequality, whereas instances where such populations are infected later in the outbreak, the inequalities have been overlooked, due to a general lack of interest in reduced risk in minority groups. It is important to note that this analysis is based entirely on the London school system. In other settings, variation between parts of the network may increase, particularly in settings that contain suburban and rural regions, where factors influencing school choice may differ from large urban centres.

The framework I applied was intended to be a method to parsimoniously evaluate the implications of the school network on inequalities explicitly. To maintain parsimony some important simplifications were made which are likely to impact the way the results of these analyses relate to epidemiology observed in a real outbreak.

Most importantly this framework only aims to simulate transmission in two settings, schools and households, between one particular age group in the population (school aged children. Firstly, there is likely to be a sizeable contribution to transmission between

school-aged children in other settings for example out of school activities such as sports. Moreover, these other transmission routes may also vary by socio-economic status and ethnic group creating further variation in out-of-school contact rates and potentially increasing rates of contact between schools in certain parts of the network. Secondly, although transmission of influenza has been demonstrated to be strongest amongst school-aged children, transmission in other age groups is likely to contribute in a meaningful way to dynamics. This also may create additional routes between schools in the network, not accounted for by connecting households.

I simplified within-school transmission to homogenous frequency-dependent transmission amongst all children in the school. In reality, transmission dynamics in school settings are likely to be more complex [21, 24, 25]. Firstly, transmission is unlikely to be homogenous but instead likely be higher between members of the same school year, class and gender [20, 25]. This would likely result in smaller eventual outbreak sizes within schools. This has an asymmetrical impact on the model since the final size would be affected but the risk of outbreak in the school would not be. In this consideration the potential for multiple introductions in different school years becomes more important, which cannot be investigated under the framework presented. Secondly, the assumption of frequency dependent transmission means that transmission rate is independent of school size. There is no clear suggestion that this would not be the case, however, if transmission were density dependent (transmission increased with school size) I would expect the results to be impacted such that pupils in regions with larger schools would be at higher risk of infection. In general, larger schools tend to be in more deprived areas suggesting that this assumption may increase inequalities overall.

As the exact data required to construct the network was not available to me, I inferred the number of contacts between schools using data on transfer of children between primary and secondary schools. This process carries uncertainty and if the true data were to become available it's use would be preferable. Another limitation this introduces is that the network produced is a bi-partite network of primary and secondary schools. This is unrealistic as you would expect some links between primary schools and between secondary schools. This would have the effect of making the network more connected, where schools have a higher degree than the one presented here. One might then expect transmission across the network to occur more quickly (in fewer generations) and this may have the impact of reducing the initial inequalities more quickly than is simulated in this model.

The simulation framework used in the main analysis requires all schools to share a single value of R_0. To evaluate the impact of this simplification I performed a sensitivity analysis allowing variation in R_0 between schools using a slight variation on the framework over fewer iterations. Although there was a small difference in degree distribution, there were not qualitative differences in the results and therefore it is unlikely that this constraint of the model has impacted the conclusions of this study overall. The sensitivity analysis evaluates the impact of normally and lognormally distributed R_0 amongst all schools as well as setting R_0 to different values in Primary and Secondary schools. There could be other systematic differences in R_0 that I have not tested which could impact inequalities, in particular, if there are differences in R_0 that correlate with deprivation status or ethnic group. To my knowledge, there is not yet any evidence of such systematic variation in R_0 between schools.

The nature of transmission between household members is unclear. The most recent findings suggest that transmission is neither frequency- nor density-dependent, but behaves somewhere between the two [31]. I chose to use a density-dependent transmission assumption for methodological convenience and have chosen a transmission probability based on a recent study of household transmission in Japan. Making a frequency-dependent assumption instead would reduce the probability of transmission between each pair of children in larger households. The role of household size would therefore be reduced in the relative connectivity of the network. My analysis of the sensitivity of the inequalities to the within-household transmission probability shows that although the overall relative risks can change significantly, the changes are not substantial enough to impact the conclusions of this analysis.

An important limitation of this model is that only state-sponsored schools have been included in the analysis. There are therefore a number of independent schools in London, which receive no money from the state and are not included in the school census. Exclusion of these schools may impact the dynamics of transmission. The full implications of this omission are not clear. Independent schools are more frequently attended by children from more affluent backgrounds and ethnic minority groups are underrepresented. On one hand, the addition of more schools in the network would increase the rate of contact between schools in more affluent areas. On the other hand, since independent schools are generally smaller than state sponsored schools, it is likely that their connection to the network is generally weaker, since there are fewer opportunities for contact pairs to form through households, therefore reducing the relative connectedness of schools in areas of higher affluence. It is also likely that independent schools are particularly well connected to each other forming strong communities of their

own, this would likely serve to partition more affluent students from the rest of the network even more than is currently observed. If this data were to become available, re-analysis of the network including independent schools may bring to light additional inequalities, which are not identified here.

The findings in of this analysis complement those of chapter 1. In similarity, they both identify that differences in transmission between populations may have the potential to promote inequalities in infection for pathogens with low transmissibility. Building on the findings of chapter 1 this analysis demonstrates that differences in overall transmission rate may not be necessary to create inequalities and that these may be a result of differences within the transmission network structure, without increase in individual rates of transmission.

To conclude, variation in the local structure of the between school transmission network is unlikely to, in itself, introduce systematic inequalities in influenza incidence over an entire uninterrupted outbreak, or when considered over multiple outbreaks. Factors affecting the containability of an outbreak (proportion of a school infected before infecting adjacent schools, and average number of schools infected by a school) may mean that some populations are more likely to observe sustained and uncontrolled outbreaks than others, however this is difficult to quantify. Due to clustering of ethnic and socio-economic groups within particular parts of the network, inequalities are likely to be common at the beginning of an outbreak. No particular group is substantially more likely to experience disproportionately high risk than others, but South Asian populations are likely to experience the most pronounced increase in risk due to their relative isolation in the network. This means that early in an outbreak, increased risk in South Asian

populations may be more likely to be reported, even when concurrent outbreaks exist in different parts of the network. There is a chance that inequalities are more frequently reported explicitly when marginalised or minority groups are disproportionately affected giving the impression that ethnic minorities and deprived communities are at substantially higher risk of disease.

6.5 References

1. Inglis NJ, Bagnall H, Janmohamed K, Suleman S, Awofisayo A, De Souza V, et al. **Measuring the effect of influenza A(H1N1)pdm09: the epidemiological experience in the West Midlands, England during the "containment" phase.** *Epidemiol Infect*. 2014, 142:428–37. doi:10.1017/S0950268813001234.

2. Balasegaram S, Ogilvie F, Glasswell A, Anderson C, Cleary V, Turbitt D, et al. **Patterns of early transmission of pandemic influenza in London - link with deprivation.** *Influenza Other Respi Viruses*. 2012, 6:e35-41. doi:10.1111/j.1750-2659.2011.00327.x.

3. Zhao H, Harris RJ, Ellis J, Pebody RG. **Ethnicity, deprivation and mortality due to 2009 pandemic influenza A(H1N1) in England during the 2009/2010 pandemic and the first post-pandemic season.** *Epidemiol Infect*. 2015, 143:3375–83. doi:10.1017/S0950268815000576.

4. Navaranjan D, Rosella LC, Kwong JC, Campitelli M, Crowcroft N. **Ethnic disparities in acquiring 2009 pandemic H1N1 influenza: a case–control study**. *BMC Public Health*. 2014, 14:214. doi:10.1186/1471-2458-14-214.

5. Wilson N, Barnard LT, Summers JA, Shanks GD, Baker MG. **Differential Mortality Rates by Ethnicity in 3 Influenza Pandemics Over a Century, New Zealand**. *Emerg Infect Dis*. 2012, 18:71–7. doi:10.3201/eid1801.110035.

6. Quinn SC, Kumar S, Freimuth VS, Musa D, Casteneda-Angarita N, Kidwell K. **Racial Disparities in Exposure, Susceptibility, and Access to Health Care in the US H1N1 Influenza Pandemic**. *Am J Public Health*. 2011, 101:285–93. doi:10.2105/AJPH.2009.188029.

7. Haroon SMM, Barbosa GP, Saunders PJ. **The determinants of health-seeking behaviour during the A/H1N1 influenza pandemic: an ecological study**. *J Public Health (Bangkok)*. 2011, 33:503–10. doi:10.1093/pubmed/fdr029.

8. Hawker J, Olowokure B, Sufi F, Weinberg J, Gill N, Wilson RC. **Social deprivation and hospital admission for respiratory infection:**. *Respir Med*. 2003, 97:1219–24. doi:10.1016/S0954-

6111(03)00252-X.

9. Mayoral JM, Alonso J, Garín O, Herrador Z, Astray J, Baricot M, et al. **Social factors related to the clinical severity of influenza cases in Spain during the A (H1N1) 2009 virus pandemic**. *BMC Public Health*. 2013, 13:118. doi:10.1186/1471-2458-13-118.

10. Munday JD, van Hoek AJ, Edmunds WJ, Atkins KE. **Quantifying the impact of social groups and vaccination on inequalities in infectious diseases using a mathematical model**. *BMC Med*. 2018, 16:162. doi:10.1186/s12916-018-1152-1.

11. Salathé M, Jones JH. **Dynamics and control of diseases in networks with community structure**. *PLoS Comput Biol*. 2010, 6:e1000736. doi:10.1371/journal.pcbi.1000736.

12. Volz EM, Miller JC, Galvani A, Ancel Meyers L. **Effects of heterogeneous and clustered contact patterns on infectious disease dynamics.** 2011, 7:e1002042. doi:10.1371/journal.pcbi.1002042.

13. Fraser C, Riley S, Anderson RM, Ferguson NM. **Factors that make an infectious disease outbreak controllable**. *Proc Natl Acad Sci*. 2004, 101:6146–51. doi:10.1073/pnas.0307506101.

14. Balcan D, Colizza V, Goncalves B, Hu H, Ramasco JJ, Vespignani A. **Multiscale mobility networks and the spatial spreading of infectious diseases**. *Proc Natl Acad Sci*. 2009, 106:21484–9. doi:10.1073/pnas.0906910106.

15. Currarini S, Matheson J, Vega Redondo F. **A Simple Model of Homophily in Social Networks Department of Economics A Simple Model of Homophily in Social Networks**. 2016. https://www.le.ac.uk/economics/research/RePEc/lec/leecon/dp16-05.pdf?uol_r=d307e306. Accessed 12 Apr 2018.

16. Eubank S, Guclu H, Kumar VSA, Marathe M V. **Modelling disease outbreaks in realistic urban social networks**. 2004, 429 May:180–4.

17. Baguelin M, Flasche S, Camacho A, Demiris N, Miller E, Edmunds WJ. **Assessing Optimal Target Populations for Influenza Vaccination Programmes: An Evidence Synthesis and Modelling Study**. *PLoS Med*. 2013, 10:e1001527. doi:10.1371/journal.pmed.1001527.

18. Worby CJ, Chaves SS, Wallinga J, Lipsitch M, Finelli L, Goldstein E. **On the relative role of different age groups in influenza epidemics.** *Epidemics*. 2015, 13:10–6. doi:10.1016/j.epidem.2015.04.003.

19. Melegaro A, Jit M, Gay N, Zagheni E, Edmunds WJ. **What types of contacts are important for the spread of infections?: using contact survey data to explore European mixing patterns.** *Epidemics*. 2011, 3:143–51. doi:10.1016/j.epidem.2011.04.001.

20. Kucharski AJ, Conlan AJK, Eames KTD. **School's Out: Seasonal Variation in the Movement Patterns of School Children**. *PLoS One*. 2015, 10:e0128070. doi:10.1371/journal.pone.0128070.

21. Conlan AJK, Eames KTD, Gage JA, von Kirchbach JC, Ross J V, Saenz RA, et al. **Measuring social networks in British primary schools through scientific engagement.** *Proc Biol Sci*. 2011, 278:1467–75. doi:10.1098/rspb.2010.1807.

22. Hens N, Goeyvaerts N, Aerts M, Shkedy Z, Van Damme P, Beutels P. **Mining social mixing patterns for infectious disease models based on a two-day population survey in Belgium.** *BMC Infect Dis*. 2009, 9:5. doi:10.1186/1471-2334-9-5.

23. Stehlé J, Voirin N, Barrat A, Cattuto C, Isella L, Pinton J-F, et al. **High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School**. *PLoS One*. 2011, 6:e23176.

doi:10.1371/journal.pone.0023176.

24. Kucharski AJ, Wenham C, Brownlee P, Racon L, Widmer N, Eames KTD, et al. **Structure and consistency of self-reported social contact networks in British secondary schools**. *PLoS One*. 2018, 13:e0200090. doi:10.1371/journal.pone.0200090.

25. Guclu H, Read J, Vukotich CJ, Galloway DD, Gao H, Rainey JJ, et al. **Social Contact Networks and Mixing among Students in K-12 Schools in Pittsburgh, PA**. 2016. doi:10.1371/journal.pone.0151139.

26. Authority GL. **London Schools Atlas**. 2019. https://data.london.gov.uk/dataset/london-schools-atlas.

27. Office for National Statistics. **2011 Census aggregate data. UK Data Service (Edition: June 2016).**

28. Biggerstaff M, Cauchemez S, Reed C, Gambhir M, Finelli L. **Estimates of the reproduction number for seasonal, pandemic, and zoonotic influenza: a systematic review of the literature**. 2014. doi:10.1186/1471-2334-14-480.

29. Furushima D, Kawano S, Ohno Y, Kakehashi M. **Estimation of the Basic Reproduction Number of Novel Influenza A (H1N1) pdm09 in Elementary Schools Using the SIR Model**. *Open Nurs J*. 2017, 11:64–72. doi:10.2174/1874434601711010064.

30. Chen S-C, Liao C-M. **Modelling control measures to reduce the impact of pandemic influenza among schoolchildren.** *Epidemiol Infect*. 2008, 136:1035–45. doi:10.1017/S0950268807009284.

31. Endo A, Uchida M, Kucharski AJ, Funk S. **Fine-scale family structure shapes influenza transmission risk in households: insights from a study of primary school students in Matsumoto city, 2014/15.** *bioRxiv*. 2019, :527507. doi:10.1101/527507.

7 Analysis D (part 1): Analysis of a between school contact network – Clustering of children by faith denomination

Objective: *Analyse the impact of faith schools on clustering of children who are susceptible to measles and resultant measles epidemiology in the Netherlands*

7.1 **Introduction**

In the previous chapter I evaluated the implications of clustering of social groups in a network of schools and differences in local network properties, for inequalities in infectious disease. I showed that for the school network in London, the greatest heterogeneities between social groups are seen for pathogens with a low reproduction number ($R_0 < 1.3$). In the case of Analysis C, a naïve population was exposed to a pathogen with low transmissibility. Another case where these dynamics may be important is in a situation of sub-optimal vaccination of a pathogen with a higher reproduction number. Moreover, where clustering of unvaccinated children within particular schools can be identified, the network may be able to identify key areas of the network (groups of schools) within which there is a particularly high risk of outbreak.

In Analysis D, which comprises this chapter and the next, I look at the school network in a setting known to have sub-optimal vaccination uptake, and where vaccination uptake has already been estimated at a school level. I propose that a school network and vaccine uptake at school level can be used to create spatial prediction of outbreaks. To test this, I constructed a school network of the Netherlands from Dutch government records and combine this with vaccination estimates at a school level. In the first part of the chapter I evaluate network structure with a particular focus on religious affiliation (which is known to associate strongly with vaccine uptake). In the second part of this analysis, described in the Chapter 8, I used the network to simulate outbreaks. To establish the importance of: A) the specific network links between schools, and B) the school-level clustering of unvaccinated children, I evaluated the spatially aggregated risk predicted by the simulations against outbreak data from a recent measles outbreak and compared to a spatial approximation of interaction between schools.

Measles in the Netherlands

A measles vaccine has been included on the routine vaccination register in the Netherlands since 1979[1]. The Mumps, Measles and Rubella (MMR) vaccine was introduced in 1989. Since introduction there have been outbreaks reported at irregular periods. In the past 20 years there have been 3 outbreaks (1999/2000, 2008 and 2013/14). The 2008 outbreak was relatively small, with 99 reported cases [2]. However, outbreaks in 1999/2000 and 2013/2014 were larger with over 3,200 and 2,700 cases reported respectively [3].

Uptake of MMR

Uptake of MMR is relatively high in the Netherlands overall with over 95% vaccinated with one dose before the age of 14 months for the past 20 years and 93% vaccinated with 2 doses by the age of 10 for the past 10 years [1]. However, uptake of MMR is highly heterogeneous across the population. As a result some municipalities report uptake as low as 66% [4] (Figure 7.1).

There is strong evidence that low uptake is related to particular socio-religious groups who refuse vaccination on mass for religious or philosophical reasons. Most notably the orthodox protestant community, approximately 1% of the Dutch population[5–8]. This group is distributed relatively sparsely along a diagonal strip from the North East to the South West of the Netherlands, and are visible in the choropleth of regional vaccination uptake rates in figure 5.1. A smaller group called the Anthroposophic community is also thought to have particularly low uptake of MMR[9, 10]. Although this community represents a smaller proportion of the Dutch population, it may still be important for

clustering of susceptible children demonstrated by an outbreak within this community in 2008 consisting of 99 reported cases.

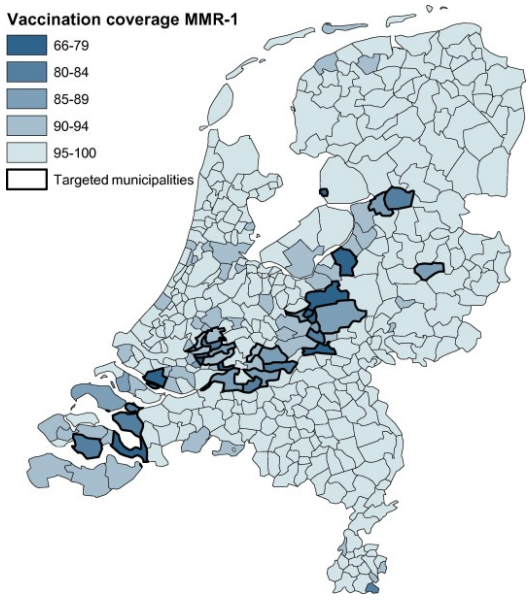


Figure 7.1 MMR first dose uptake by 14 months by municipality in the Netherlands (Lochlainn et. al., 2017) [4]

Motivation

It has been suggested that clustering of unvaccinated people by socio-religious groups may be responsible for the potential for very large outbreaks to occur, such as in 1999/2000 and 2013/14. Periodic outbreaks have meant the age of most individuals infected with measles has been lower than the rest of Europe despite high vaccine uptake. For example in the large outbreak in 2013/14, the vast majority (90%) of cases occur in school-aged children [11]. By comparison, the proportion of cases in this age group across the whole of Europe each calendar year varied between 50 and 80% in the decade between 2008 and 2018 [12]. From chapter 4 and previous analysis [13], a great deal of potentially infectious contact between children is likely to occur within the social networks

associated with their school and household. On this basis, I propose that the structure of the Netherlands education system might provide a means of quantifying the extent of social clustering within particular groups known to refuse vaccination at high rates.

Network analysis is a broad discipline for studying complex systems. One example of its use is the evaluation of networks of patient transfer between health care units to evaluate the risk of transmission of antimicrobial resistant infections across healthcare systems in the United Kingdom and the Netherlands[14–17].

In a similar way, for this analysis I constructed a network of schools in the Netherlands, I analysed key properties of the network pertaining to clustering of unvaccinated children in schools using various network analysis techniques, to identifed whether the structure of the school network may contribute to clustering of unvaccinated children at local and national scales.

Overview of Education in the Netherlands

General structure

Although education is compulsory in the Netherlands from the age of 5 to the age of 16, the majority of children start attending from the age of 4 and some elements of high school can continue until the age of 18. Students attend a primary school (basisschool: ‘basic school’) until the age of 12. Following this they are transferred into secondary school (voortgezet onderwijs: "continued education"). Secondary school is divided into three tiers (Figure 7.2): pre-vocational secondary education for 4 years (VMBO), senior general secondary education for 5 years (HAVO) or pre-university education for 6 years (VWO) [18]. Each of these prepares students for different types of further education or careers.

Notably there are many instances where there are multiple streams represented within a single institution, many of them on the same site.

Advanced or struggling students can also move to a more appropriate tier as they pass through secondary school.

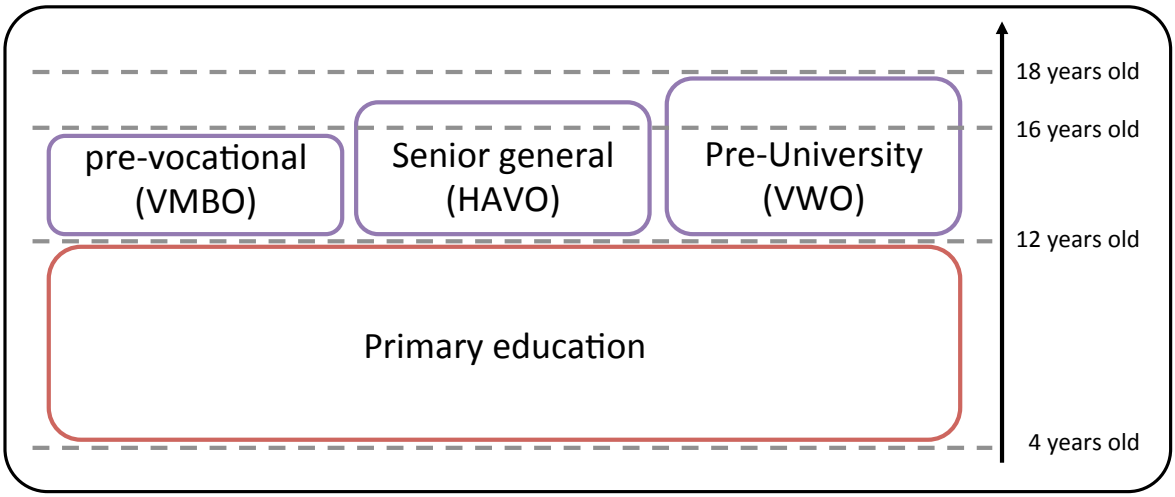


Figure 7.2 The education system in the Netherlands has 2 stages. The second stage has 3 tiers based on academic ability.

Special schools (bijzondere)

Another important discrimination within the school system is by faith or educational philosophy. In particular, bijzondere schools are government-funded institutions, but affiliated with either a religious community or educational philosophy. There are 27 different denominations of schools coded for the analysis in this chapter (table 7.1): public schools (Openbaar), a denomination to represent all non-religious special denominations (Algemeen bijzonder); 15 religious denominations and 9 collaborations between the existing 17 coded denominations.

Special schools comprise a major part (~69%) of the Dutch education system, with non-religious public (Openbaar) schools only comprising ~31% of schools and ~28% of pupils. Approximately 9% of special schools are non-religious (Algemeen bijzonder).

Denomination	*Dutch name*	*Schools*	*Primary*	*Secondary*
Public school	Openbaar	2810	2466	344
Roman Catholic	Rooms-Katholiek	2554	2258	296
Mainstream protestant	Protestants-Christelijk	2147	1904	243
Special educational philosophy	Algemeen bijzonder	807	579	228
Dutch Reformed	Reformatorisch	208	181	27
Reformed liberated	Gereformeerd vrijgemaakt	118	118	0
Anthroposophic	Antroposofisch	81	70	11
Islamic	Islamitisch	44	43	1
Interconfessional	Interconfessioneel	21	15	6
Reformed Liberated	Gereformeerd	18	0	18
Evangelical	Evangelisch	16	12	4
Hindu	Hindoeïstisch	6	6	0
Other	Overige	4	0	4
Jewish	Joods	2	2	0
Moravian Church	Evangelische broedergemeenscha	2	2	0
Potestiant/Evangelical	Protestants-Christelijk/Evange	1	0	1
Jewish Orthodox	Joods orthodox	1	0	1
Potestiant/Reformed	Protestants-Christelijk/Reform	1	0	1

Table 7.1 Faith schools in the Netherlands. The number of schools, primary schools and secondary schools in each faith denomination in the Netherlands.

The remaining denominations are religious and comprise 60% of schools in the Netherlands. Of these denominations Roman Catholic (Rooms-Katholiek) and mainstream Protestant (Protestants-Christelijk) schools are the distinctively the largest

denominations by proportion of schools (~28% and ~23% respectively) and proportion of students attending (~30% and 21% respectively). The remaining 9% of schools are made up of 12 relatively small denominations, the three largest of which are Dutch Reformed (Reformatisch), Reformed-Liberated (Gereformeerd vrijgemaakt) and Anthroposophic (Antroposophisch). It is these three denominations, which most closely align with populations that are known to refuse vaccination for religious reasons.

7.2 Methods

To evaluate whether schools cluster by faith denomination I constructed a network of schools based on links through households and performed a series of descriptive analyses of this network. Firstly, to identify key predictors of community structure without explicitly testing specific faith affiliations I partitioned the network to establish its natural community structure. I then analysed this structure to evaluate the importance of belonging to geographic administrative regions and religious faith denominations for belonging to particular communities of schools. Secondly, I explicitly analysed the connections between schools of the same and different faith dominations calculating homophily by faith denomination and then estimated long range connections by evaluating the relationship between geographic distance and shortest paths on the school network.

School data

School and pupil data were provided by the Dutch ministry for education (DUO), which holds data on each school (n. 9200) and individual child in the educational system on the 31st October 2013. This data includes place of residence and school attended. DUO

publishes aggregates of this data by school including educational stage (e.g. primary or secondary), location by postcode and geographic coordinates, residence of students aggregated at the 4-digit postcode (PC4) level, and religious affiliation. DUO also keeps detailed data on pupils, including household level residence data. I was able to collaborate with members of the data team to access bespoke aggregates pertinent to the analysis discussed below.

School contact network

With the data provided by the Dutch Education Executive Agency (DUO) I constructed a network of schools where edges were weighted by the number of unique contact pairs, where a contact pair comprises two children who reside in the same household but attend different schools. This is discussed in detail in Chapter 5. Networks were constructed and analysed using the NetworkX package in the python programming language [19, 20].

Analysis of school network characteristics – general understanding of the network

Analysis of degree distributions

I assessed the general properties of the network starting with the degree (number of connected schools) and weighted degree (number of unique pairs) of each school. I evaluated the distributions of these measures and relationship between weighted degree and degree to quantify a distribution of the mean number of pairs per school. I repeated this analysis for primary and secondary schools separately and individually for schools of Roman Catholic, Mainstream Protestant, Dutch Reformed and Anthroposophic denominations to assess any differences in these key characteristics by denomination.

Finally, to establish the role of primary and secondary schools in the network I calculated the proportion of contact pairs between two primary schools, two secondary schools and a primary and secondary school.

Communities within the network

Before explicitly investigating the strength of connection between particular denominations, I evaluated natural communities of schools within the network. The aim of this is to reveal groups of schools that are well connected with a hypothesis-free methodology.

Community detection algorithms offer a set of tools for evaluating the structure of a network in a way that is naïve to the node labels (e.g. denomination of the school). Typically, the communities identified represent some modular structure[21, 22], which aligns with the particular constraints defined in the framework used to detect it.

In previous, similar evaluation of network communities, networks with a strong geographical component have demonstrated community structures which broadly form geographically contiguous clusters[14]. If school choice were entirely defined by geographical proximity, the same may be expected for the school network. The presence of religious affiliation with schools presents an additional factor: should the religious affiliation of schools be substantial, then communities may be expected to reflect these groups as well as geographical groups, which would indicate strong connectivity between these schools on the basis of shared faith. In turn, if that faith is associated with low vaccination, high connectivity arising as a result of religious affiliation may correspond

to high connectedness between schools with low MMR uptake. vaccination, may correspond to high connectedness between schools with low MMR uptake.

Community detection framework

I used a variation of the Leiden[23] to partition the graph, which modifies the popular Louvain algorithm to overcome a key problem of identifying disconnected communities (communities which from multiple components). I chose this framework as the community definition is explicit in this method, which makes interpretation of the resulting communities more straightforward in this case as I explicitly sought assortative communities, which is problematic for, generally more robust, frameworks based on statistical inference. The version of the Leiden algorithm I used maximises a quality function:

$$Q = \sum_{ij} \left(A_{ij} - \frac{\gamma k_i k_j}{2m} \right) \delta(C_i, C_j)$$

where A is the adjacency matrix, k_i is the (weighted) degree of node i, m is the total number of edges (or total edge weight), C_i denotes the community of node i and $\delta(C_i, C$) 1 if $C_i = C_j$ and 0 otherwise.

γ is a resolution parameter taking a value between 0 and 1, which moderates the scale of the communities detected: the higher the value of γ, the smaller the communities detected. To establish the most meaningful scale of communities, I partitioned the network with values of γ between 0.1 and 1. I evaluated the partitions against four metrics

to establish the most appropriate resolution parameter. For each of the metrics higher values suggest a better partition for our purposes. The metrics I used were:

Internal edge density[24]*:*

This metric measures the proportion of possible edges that are present in community C, expressed as:

$$\rho_C = \frac{m_C}{½\, n_C(n_C - 1)}$$

where m_C is the number of edges internal to community C and n_C is the number of schools in community C. This takes a value between 0 and 1 and provides a quantification of the absolute connectivity within the community, irrespective of connectivity with other communities.

Modularity density [25]*:*

Modularity density normalises the quality function by the number of schools in the community, hence removing dependence on community size. This provides a better comparison of modularity between partitions with different community sizes. This is expressed:

$$Q_{dens}(S) = \sum_{C \in S} \frac{1}{n_C} \left(\sum_{i \in C} k_{iC}^{in} - \sum_{i \in C} k_{iC}^{out} \right)$$

where n_C is the number of schools in C, k_{iC}^{in} is the degree of node i within C (edges to schools inside the community) and k_{iC}^{out} is the degree of node i outside C (edges to schools outside the community) for a partition S.

Neman Girvan modularity [26]*:*

A classic metric of how strongly defined communities in a partition are from the rest of the graph. Is calculated as:

$$Q_{NG}(S) = \frac{1}{m}\sum_{C \in S}\left(m_C - \frac{(2m_C + l_C)^2}{4m}\right)$$

where m is the number of graph edges, m_C is the number of community's edges, l_C is the number of edges from schools in C to schools outside C.

Surprise [27]*:*

Surprise is a quality metric assuming that edges between nodes emerge according to a hyper-geometric distribution. According to the Surprise metric, the higher the score, the less likely that the communities detected occurred at random and therefore the better the quality of the partition.

Modularity optimisation algorithms, such as the Leiden algorithm used here, are stochastic processes, and hence the exact partition recovered in each iteration of the algorithm depends on the order in which the nodes are sorted[21, 28–30]. First, I explored the variation in partitions by evaluating the normalised mutual information between each pair of partitions.

Ensemble partition

To establish a stable summary partition, I ensembled the partitions resultant from multiple iterations to find a consensus partition of the network. For this I employed a method of ensembling proposed by Lancichinetti and Fortunato[29], following the steps:

1. Find N_c partitions of the graph using the community detection algorithm
2. Construct a similarity matrix D where each entry D_{ij} is the proportion of partitions that nodes i and j are partitioned into the same community.
3. Find N_c partitions of the graph with adjacency matrix defined by the similarity matrix.
4. Perform steps 2 and 3 until all N_c partitions are identical.

Although this method provides a stable partition, it does not necessarily produce the optimal partition in terms of modularity or statistical significance[21, 29]. Therefore, in addition to the consensus partition I identified schools that are partitioned into the same communities 100% of iterations. I also evaluated the particular partition with the highest Newman Girvan modularity[26], as this represents the initial partition with the most clearly defined communities.

Analysing community composition

Finally, to quantify the composition of the communities in the partition in terms of denomination and administrative province, I calculated the mean pairwise probability that any two schools of the same particular denomination or province fall into the same community over the partitioned networks. Explicitly, for each denomination, I calculated

proportion of times each school in the denomination was in the same community as each other school in the denomination. I then calculated the mean pairwise probability as the mean of these values over all schools. I then repeated the same for each province.

Preferential contact between schools of the same denomination

To assess the local connectivity of schools by denomination, I used two network measures to evaluate if schools of certain denominations are more connected to schools of the same than would be expected at random.

Firstly, I quantified the preference for connections between schools of the same religious affiliation by calculating the Basic Homophily (H_i) by denomination for the 17 coded individual denomination, which is the average proportion of connected schools that belong to the same denomination.

$$H_i = \frac{s_i}{s_i + d_i}$$

Here, s_i is the number of neighbours of the same denomination and d_i is the number of neighbours of a different denomination. Secondly, I calculated the Coleman Homophily Index[31], which gives the proportion of neighbouring schools that align with the same religious denomination relative to the proportion of schools that belong to that denomination.

$$IH_i = \frac{H_i - w_i}{1 - w_i}$$

To address longer-range connections in the network, I compared geographical and network distance between pairs of schools, similarly to Donker et al [16]. To ensure network distance was shortest for the most connected schools, I defined it as the length of the shortest path between schools on the reciprocal contact network. The edges in the reciprocal network are equal to the reciprocal of the number of unique contact pairs between pairs of schools. The network distance for a path is therefore $\sum_{i=1}^{N_{path}-1} \frac{1}{C_{i+1,i}}$ where $C_{i+1,i}$ is the number of contact pairs between consecutive schools in a shortest path and N_{path} is the number of schools in the shortest path (Figure 7.3).

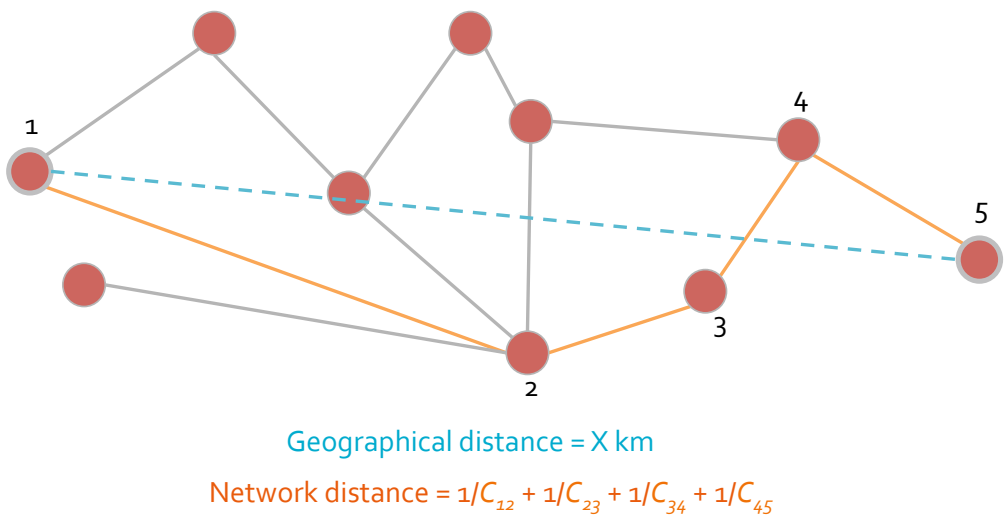


Figure 7.3 Calculation of network distance between schools 1 and 5 is the sum of the edges along the shortest path between those schools.

I calculated the network distance and geographic distance of 1000 randomly sampled pairs of schools from the biggest faith-based school denominations in the Netherlands: Roman Catholic (Rooms-Katholic) and mainstream Protestant (Protestants-Christelijk). I also calculated the distances for Dutch reformed and Anthropasophic denominations, which are most closely associated with low vaccination uptake.

For each denomination I calculated the ratio of network distance and geographic (km) distance (distance ratio) for each of the pairs sampled. Schools with a low distance ratio are more closely connected on the network relative to their geographic distance than schools with higher distance ratios. I calculated distance ratios for pairs of schools of the same denomination to that of schools in the rest of the network, defined as all schools not associated with that denomination.

Connectedness of schools through households may be substantially impacted by demography, geography and infrastructure, such as population density and location of waterways and transport links. The relationship between geographical distance and network distance may therefore be highly dependent on the locations of the schools considered. To account for the potential effect of school location, I sampled a school from the 'rest of the network' from the same two-digit postcode area as each school sampled from the particular denomination studied.

To correct for the potential for homophily within the particular denomination of interest to impact the shortest network path between schools in the 'rest of the network', I removed all schools of this denomination when calculating the network distances for the 'rest of the network'. This correction itself could introduce bias by reducing the number of schools in the network. To balance this effect when I calculated network distances between schools of a particular denomination, I removed randomly selected schools located in the same two-digit postcode area as those from the denomination of interest.

7.3 Results

Network Characteristics

I constructed a network of schools connected by pairs of students who reside in the same household (e.g. siblings) but attend different schools (contact pairs) using a national school registration dataset from the Netherlands, provided by the Education Executive Agency (DUO). The total number of schools in the network was 8095 including 6736 primary schools and 1359 secondary schools. The network is connected i.e. there exists a path from each school to every other school. Secondary schools have a mean degree of 103, (schools connected by at least one contact pair, the number of edges that connect each school to the network); this is more than primary schools, which have an average degree of 22 (Figure 7.4). The same is reflected in the weighted degree (mean number of contact pairs with all connected schools i.e. the sum of the edge weights), where secondary schools have a median of 660 and primary schools have a median of 88. The major special schools (Roman Catholic and mainstream Protestant) exhibited similar median connected schools, 97 and 101 respectively for secondary schools and 21 for both denominations in primary. The median number of unique contact pairs was slightly higher with values of 752 and 672 for Roman Catholic and Protestant schools respectively.

Of the denominations most associated with low MMR uptake, Dutch reformed secondary schools had a much higher median number of unique contact pairs 1553, this however was paired with a lower median value for connected schools than the rest of the network, 85. Primary schools of this denomination had a similar median number of connected schools to the rest of the network 23. However, like secondary schools, the median number of unique contact pairs was much higher at 230.

Primary and secondary Anthroposophic schools both had a higher median number of connected schools, 149 and 26 respectively. However, the median number of unique contact pairs was lower than that of the whole network, 574 for secondary and 81 for primary.

The relationship between weighted and un-weighted degrees shows that in general the number of contact pairs per school is highest for those of the Dutch Reformed denomination. In contrast the number of contact pairs per connected school is generally lower in Anthroposophic schools than in the rest of the network.

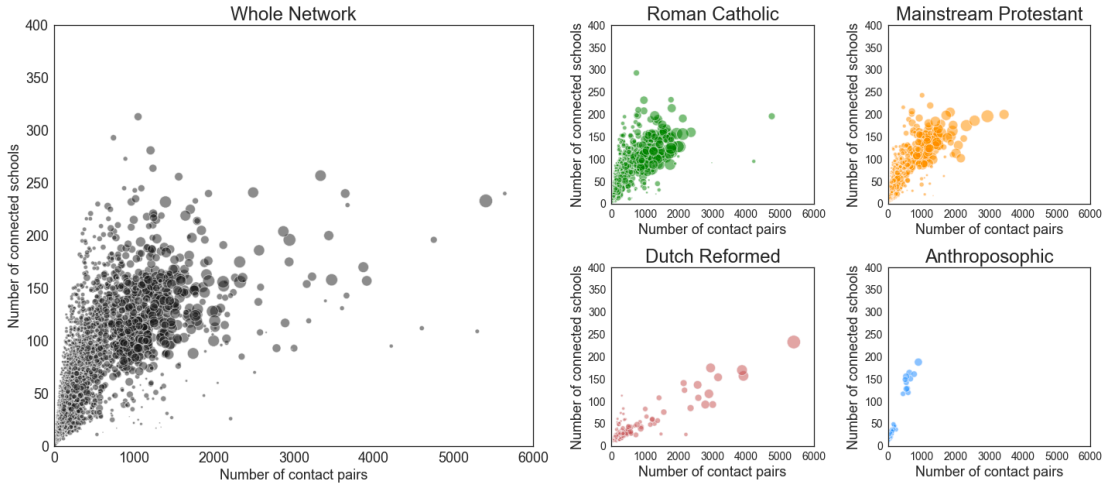


Figure 7.4 Scatter plots of degree (number of connected schools) and weighted degree (the number of unique pairs). Points show schools from A) the whole network, B) roman catholic denomination, C) Mainstream protestant denomination, D) Dutch Reformed denomination and E) Anthroposophic denomination. Marker size indicates school population size.

Communities in the Network

I partitioned the network using a range of resolution parameters γ between 0.1 and 1.0 to define the quality function, all metrics increased with resolution parameter (Figure 7.5). This suggests that smaller communities had both clearer definition from the rest of the

network and were less likely to occur at random than larger communities found using the lower values of γ. For this reason, I chose to use a resolution parameter of $\gamma = 1.0$ for the remainder of the analysis, which corresponds to the unmodified Leiden algorithm [23].

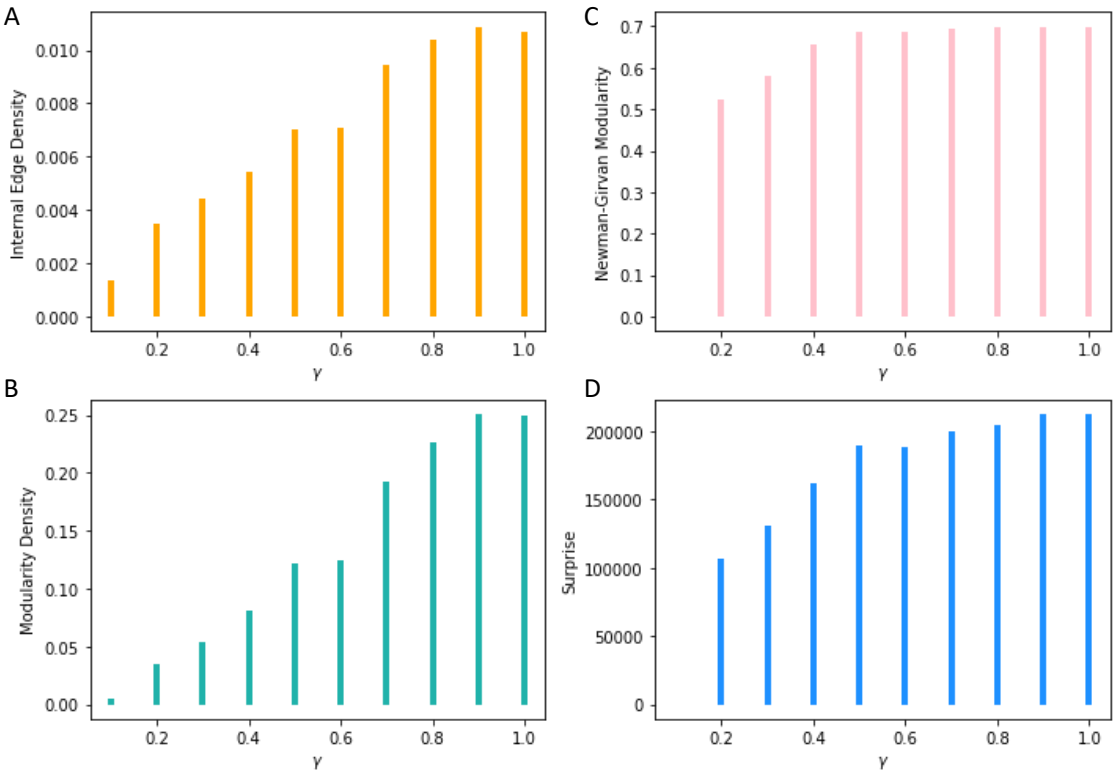


Figure 7.5 Quality metrics for various values of resolution parameter γ for partitions recovered using the modified Leiden algorithm.

Panels A to D show the scores for Internal Edge Density, Newman-Girvan Modularity, Modularity Density and Surprise respectively.

I generated 20 partitions of the network using the Leiden algorithm (Figure 7.6 A). Each of these had a clear geographical component to the community structure. Although each of the partitions was unique, the normalised mutual information was high between all partitions (greater than 0.8 and greater than 0.9 in most cases) (Figure 7.6 D). This

indicates that the partitions generated in each iteration were similar, which in turn corresponds with relatively stable community structure in the network.

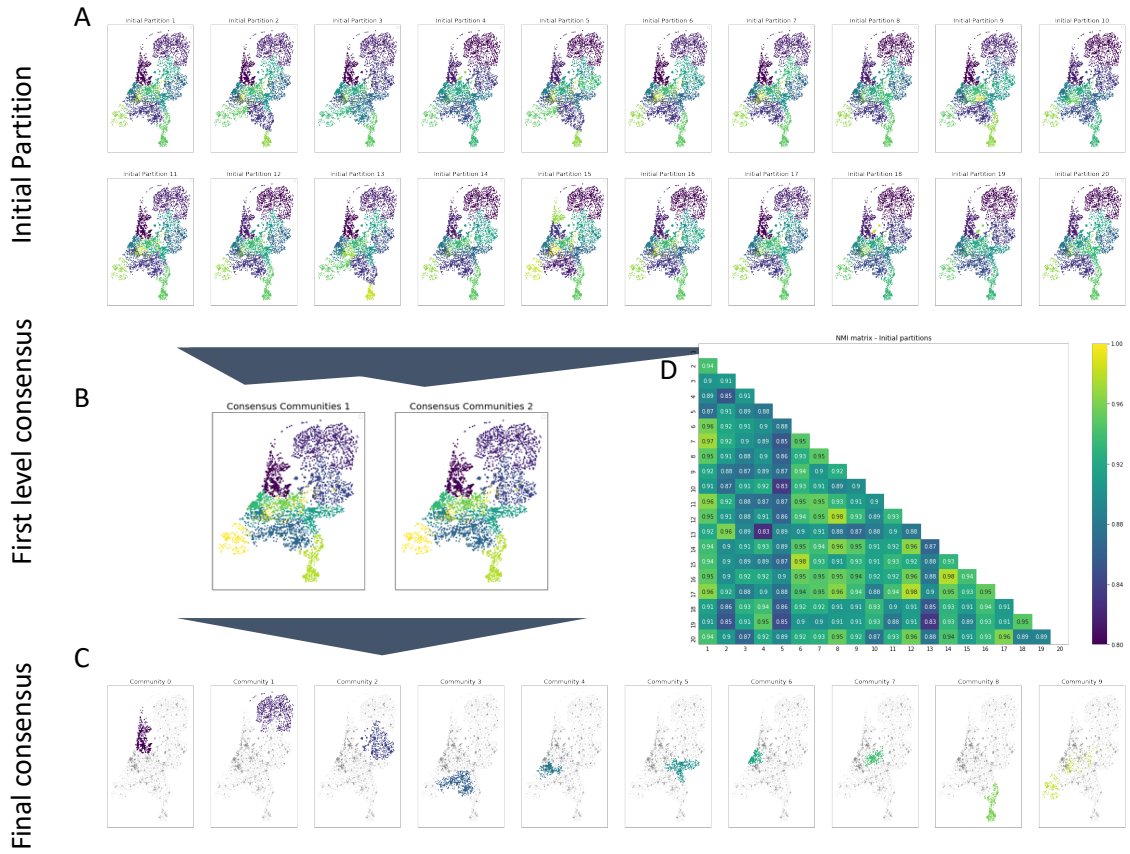


Figure 7.6 partitions of the school network in the Netherlands

A to C show the locations of schools in the Netherlands, the colour of the markers indicates the community of schools in the partition. The panels in A show the 20 initial partitions, B shows the partitions from the first round of the ensembling algorithm. C shows each community in a separate panel in the final consensus partition, grey points show the locations of other schools (not in the community in that panel). D shows a matrix normalised mutual information (NMI) between the initial partitions.

Using the 20 partitions, the ensemble algorithm converged after two iterations, attesting to the stability of the initial partitions. Further, the first round revealed two unique

partitions which were very similar (Figure 7.6 B), with a normalised mutual information score of 0.98 between them.

The consensus partition itself revealed largely geographically organised communities (Figure 7.6 C) with high probability of schools in the same province being assigned the same community. In general, any preference of connection between schools of the same religious affiliation was not sufficient to overpower the strong geographical component in the communities.

However, one community (Figure 7.6C, community 9) was not comprised only of schools that were close geographically. Like the other communities in the consensus partition, this community, labelled as community 9, had a strong geographical component containing an overwhelming majority of schools in the province of Zeeland; however, the community also contained 151 schools in other provinces, 129 of which were affiliated with the Dutch reformed denomination and 22 to the mainstream protestant denomination (Figure 7.7). For all the provinces represented in the community, more than 80% of Dutch reformed schools were included in the community. In each province other than Zeeland, fewer than 2.5% of mainstream protestant schools were included in the community (Table 7.2). The initial partition with highest modularity also included a community similar to that shown in Figure 7.6 with the vast majority of schools overlapping between the two communities (Appendix D Table 1, Figure 1). Similarly grouping schools only if they were partitioned into the same community in every initial partition revealed that 107 Dutch Reformed schools from outside Zeeland remained in the same community in every partition (Appendix D Table 2, Figure 2), which form a subset of those in the community shown in Figure 7.7.

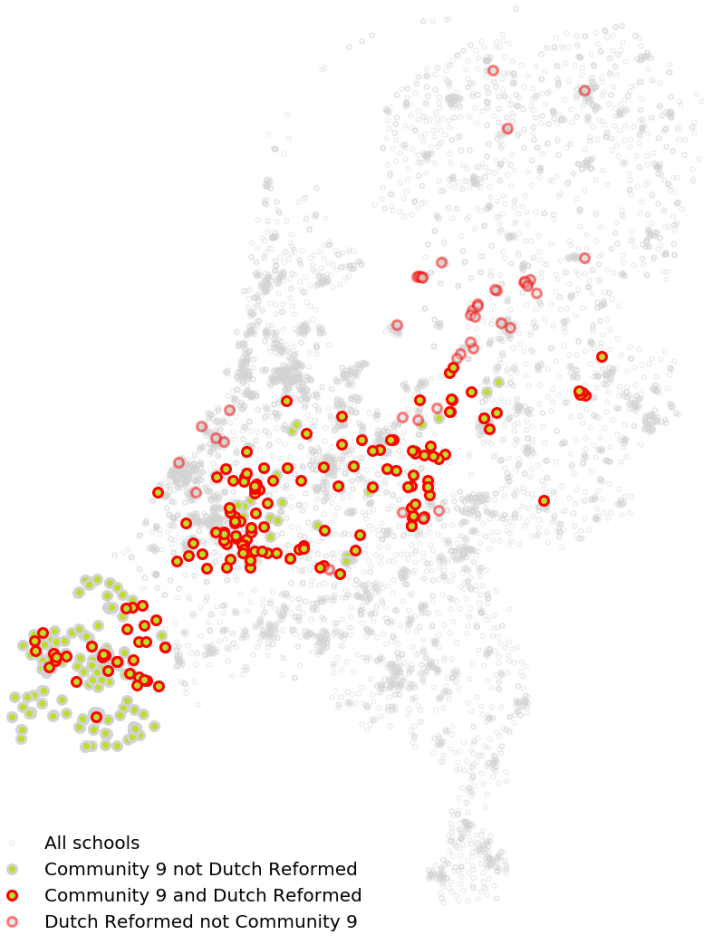


Figure 7.7 Community 9 of the consensus partition

Locations of schools in the Netherlands. Markers with green fill show schools in Community 9, markers with grey fill indicate schools not in Community 9, red edges indicate Dutch Reformed schools, and grey edges indicate schools of other denominations.

The pairwise probability that schools of the same province fell into the same partitioned communities was high with a mean of 0.75 (Table 7.3, Appendix D Figure 1 and Figure 2). In contrast the mean pairwise probability that schools of the same denomination were partitioned into the same communities was much lower with a mean of 0.28. There were three denominations with exceptionally high mean pairwise probability, they include Jewish schools (Joods) and Moravian Church (Evangelische broedergemeenscha) schools, which both have only a small number of geographically clustered schools in the

population. The third was Dutch Reformed (Reformatorisch) schools, which despite relatively sparse geographic distribution and being the third largest religious denomination in the school system, showed a mean pairwise probability of 0.55.

Province	Denomination	All schools	In Consesnsus community	%
Gelderland	Protestants-Christelijk	356	6	1.7%
	Reformatorisch	52	42	80.8%
Noord-Brabant	Reformatorisch	4	3	75.0%
Noord-Holland	Reformatorisch	2	2	100.0%
Overijssel	Reformatorisch	22	7	31.8%
Utrecht	Protestants-Christelijk	207	4	1.9%
	Reformatorisch	16	16	100.0%
Zeeland	Algemeen bijzonder	22	19	86.4%
	Antroposofisch	1	1	100.0%
	Gereformeerd vrijgemaakt	3	3	100.0%
	Openbaar	96	79	82.3%
	Overige	1	1	100.0%
	Protestants-Christelijk	63	56	88.9%
	Reformatorisch	36	36	100.0%
	Rooms-Katholiek	48	45	93.8%
	Samenwerking PC, RK	10	10	100.0%
Zuid-Holland	Protestants-Christelijk	524	12	2.3%
	Reformatorisch	65	59	90.8%

Table 7.2 Composition of Community 9 in the final consensus partition detailing number of schools by province and denomination in the community and in the whole network.

Denomination	Mean pairwise probability
Joods	0.925, (0.739, 1.000)
Reformatorisch	0.550, (0.546, 0.554)
Evangelische broedergemeenscha	0.500, (0.000, 1.000)
Hindoe□stisch	0.280, (0.081, 0.479)
Gereformeerd	0.268, (0.220, 0.317)
Interconfessioneel	0.233, (0.196, 0.271)
Gereformeerd vrijgemaakt	0.210, (0.203, 0.217)
Evangelisch	0.149, (0.105, 0.193)
Rooms-Katholiek	0.147, (0.147, 0.147)
Islamitisch	0.147, (0.132, 0.162)
Openbaar	0.129, (0.128, 0.129)
Protestants-Christelijk	0.125, (0.124, 0.125)
Algemeen bijzonder	0.122, (0.121, 0.122)
Antroposofisch	0.118, (0.110, 0.125)
Mean	**0.28**

Province	Mean pairwise probability
Groningen	0.996, (0.996, 0.996)
Friesland	0.988, (0.988, 0.989)
Noord-Holland	0.950, (0.949, 0.950)
Zeeland	0.887, (0.885, 0.889)
Limburg	0.857, (0.855, 0.858)
Noord-Brabant	0.784, (0.784, 0.785)
Drenthe	0.755, (0.753, 0.758)
Utrecht	0.723, (0.722, 0.725)
Flevoland	0.734, (0.681, 0.686)
Overijssel	0.546, (0.544, 0.547)
Zuid-Holland	0.442, (0.441, 0.442)
Gelderland	0.343, (0.342, 0.344)
Mean	**0.75**

Table 7.3 The mean pairwise probability (95% CI) that schools of each denomination and province are partitioned into the same community.

calculated from 20 iterations of the Louvain modularity maximisation algorithm. Denominations with only 1 school have been precluded.

Homophily by faith denomination

I calculated the Basic Homophily (BH) and Coleman Homophily Index (CHI) of the schools by denomination. The majority of denominations had positive homophily indices, suggesting that households are more likely to have children in two or more schools of the same denomination than would be expected at random.

The four denominations with the highest Coleman Homophily index were Dutch Reformed, Athroposophic, Roman Catholic and Mainstream Protestant, With CHI ranging from 0.62 to 0.12. Notably the two denominations with the highest CHI were Dutch Reformed (Reformatorisch) (BH = 0.63, CHI= 0.62) and Anthroposophic (Antroposofisch) (BH = 0.25, CHI = 0.24), which are the two denominations thought to most closely align with populations who systematically refuse vaccination (Figure 7.8).

Distances across the network

I compared network and geographic distances for schools within Roman Catholic, Mainstream Protestant, Dutch Reformed and Anthroposophic denominations.

The mean ratio of network to geographic distance (calculated for a sample of 500 schools) was 2.33 $x10^{-3}$ (2.05 $x10^{-3}$ - 2.62 $x10^{-3}$, 95% CI) $pairs^{-1}$ km^{-1} for the whole network, 0.96 $x10^{-3}$ (0.84 $x10^{-3}$ - 1.08 $x10^{-3}$, 95% CI) $pairs^{-1}$ km^{-1} for Dutch Reformed, 3.03 $x10^{-3}$ (2.65 $x10^{-3}$ - 3.42 $x10^{-3}$) $pairs^{-1}$ km^{-1} for Anthroposophic, 2.76 $x10^{-3}$ (1.89 $x10^{-3}$ - 3.63 $x10^{-3}$, 95% CI) $pairs^{-1}$ km^{-1} for Roman Catholic and 2.41 $x10^{-3}$ (2.15 $x10^{-3}$ - 2.67 $x10^{-3}$, 95% CI) $pairs^{-1}$ km^{-1} for mainstream Protestant (Figure 7.9). This indicates that The Dutch Reformed denomination alone forms extended chains of schools strongly linked through households, whereas the other denominations are generally as connected as any schools in the whole network.

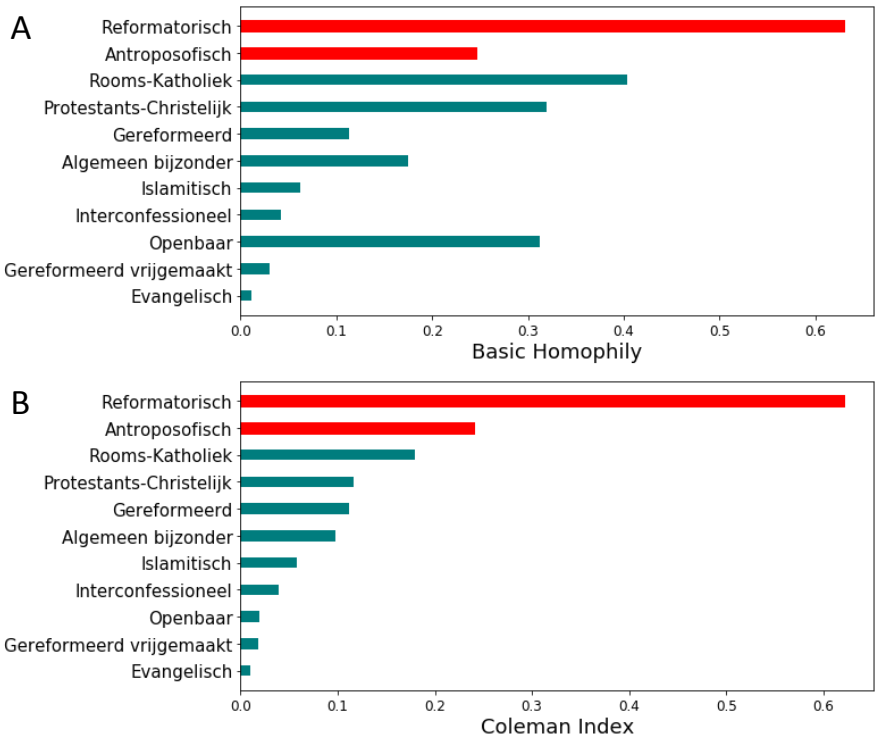


Figure 7.8 The 11 denominations with the highest Coleman Homophily Index (CHI).

A) Bars show the basic homophily of each denomination. B) Bars show the CHI of each denomination. In A and B the red bars highlight the Dutch reformed and Anthroposophic denominations, where vaccination uptake is known to be low.

I calculated the network and geographic distances for Dutch Reformed and Anthoroposophic denominations and compared them with a geographically equivalent comparator sample from the rest of the network. The scatter plots of network distance against geographic distance revealed that in both cases network paths were shorter between the Dutch Reformed and Anthroposophic schools than between randomly selected geographically equivalent schools.

The distance ratio (network distance divided by geographic distance) distribution was lower for Dutch Reformed schools and Anthroposophic schools than their comparison

samples. With mean distance ratios of 0.90×10^{-3} pairs^{-1} km^{-1} (0.77×10^{-3} - 1.04×10^{-3}, 95% CI) and 2.37×10^{-3} pairs^{-1} km^{-1} (2.07×10^{-3} - 2.66×10^{-3}, 95% CI) for Dutch reformed and Anthroposophic schools respectively and 5.70×10^{-3} pairs^{-1} km^{-1} (4.81×10^{-3} - 6.59×10^{-3}, 95% CI) and 4.40×10^{-3} pairs^{-1} km^{-1} (3.85×10^{-3} - 4.95×10^{-3}, 95% CI) for their respective comparators.

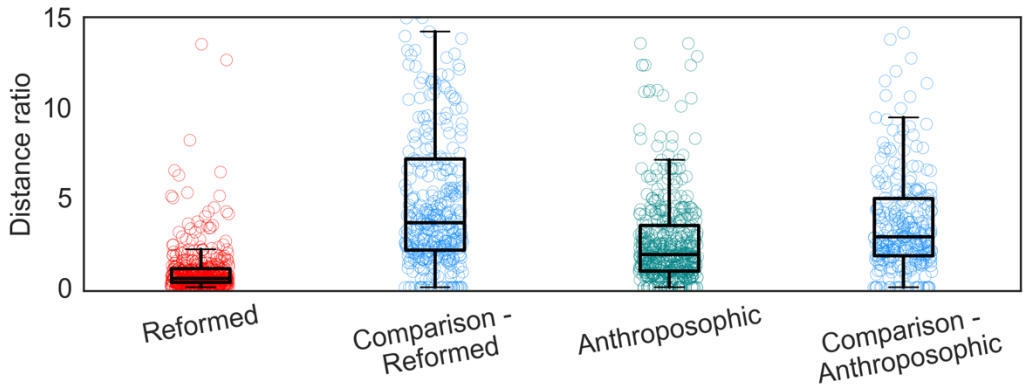


Figure 7.9 Boxplot of distance ratio for pairs of Dutch Reformed and Anthroposophic schools and geographically equivalent sample from the rest of the network.

7.4 Discussion

Due to close association of certain socio-religious groups with low MMR uptake and the tradition of faith schools in the Netherlands, it has previously been suggested that unvaccinated children cluster within schools of particular faith denomination[9, 11, 32]. However, it has been unclear how large-scale and long-range clustering of unvaccinated populations is maintained when uptake of MMR is high within the general population. I constructed a network of schools in the Netherlands from national school data, where schools are the nodes of the networks and the edges are weighted by the number of unique contact pairs, which form by residing in common households. I analysed the network to

assess the potential for local and long-range clustering of unvaccinated children by contact connections through households.

Degree and weighted degree distributions of the whole network and subsets of the network by denomination (Mainstream Protestant (Protestants-Christelijk), Roman Catholic (Rooms-Katholiek), Anthroposophic (Antroposofisch) and Dutch Reformed (Reformatorisch)) show that the largest denominations, Roman Catholic and Mainstream Protestant, have a very similar distribution to the whole network, however the denominations most associated with low vaccination uptake had distinct differences. Dutch Reformed presents a very low number of connected schools but with much higher number of unique contact pairs per school, suggesting that parents of children attending Dutch Reformed schools are more consistent and selective in where they send their children to school. This could be amplified by larger than average family sizes[33], as a larger number of school going children may result in a higher number of unique contact pairs through households. Anthroposophic schools displayed the opposite relationship, with a lower number of unique contacts but a relatively large number of connected schools, suggesting that children from Anthroposophic schools have siblings in a more diverse selection of schools. It could also indicate that children in Anthroposophic schools belong to households with a lower number of children than average.

The analysis suggests that schools are generally clustered by province, where most communities in the consensus partition were strongly clustered geographically, and not by denomination. However, one community contained a large number of Dutch Reformed schools, independent of their location, suggesting high degree of connectedness between these schools relative to the schools that were closest geographically.

Most denominations had a positive Coleman Homophily Index, indicating that in general schools are more likely to be connected to schools of the same denomination than would be expected at random. The CHI was particularly high for Dutch Reformed schools (0.62) and Anthroposophic schools (0.24), which are both associated with low vaccination uptake. High homophily between schools with low vaccination uptake indicates increased propensity for an outbreak in one school with low uptake to seed to another of the same denomination.

By evaluating the ratio of network and geographic distances, it was evident that the Dutch reformed schools had much shorter paths between them compared to the rest of the network in general. Although the paths were consistently shorter between Anthroposophic schools compared to the rest of the network, the difference was not as pronounced as that for the Dutch Reformed schools. This indicated that although *both* Dutch Reformed schools and Anthroposophic schools showed high homophily, the Dutch reformed schools formed long-range chains of strong links between schools, whereas the strong connections between Anthroposophic schools was much more local. The low values of distance ratio for Dutch Reformed schools suggest that in addition to local homophily, household links between Dutch Reformed schools form strong chains across the network independent of geographic, demographic or infrastructural effects.

The distinct network proximity of geographically distant Dutch Reformed schools indicates the potential for long range and large-scale clustering of unvaccinated children from this particular population, without the need for long-range transmission events. However for Anthroposophic schools these the connection between distant schools was

not as strong, suggesting that these schools do not contribute to long range clustering of unvaccinated children, if an outbreak occurred in a school of this denomination long-range transmission events are likely to be required to spread an outbreak beyond the area local to where the first cases occur.

This finding is consistent with outbreaks in the past decade, where an outbreak in an Anthroposophic school in 2008 was contained to 99 cases in a relatively local area, whereas an outbreak in the Dutch Reformed community in 2014 spread to distant parts of the country infecting an estimated 30,000 people, 90% of which were children and 84% belonged to the Dutch Reformed community.

There are a number of important limitations to this analysis. Firstly, the framework cannot capture connectivity between schools through any means other than household. It may be the case that connectivity through other means, such as activities connected religious provide clearer connections between other religious affiliations that are not highlighted here. Furthermore, religious identity may not in all cases strongly align with school denomination, however there is independent evidence that there is low vaccine uptake in schools affiliated with Dutch Reformed and Anthroposophic groups.

Community detection is one of the most researched areas in network science and its application is of much interest in multiple fields of scientific study. Over recent decades many algorithms have been developed to detect communities in a network, which vary greatly in approach and even the definition of the “communities” which they seek [21], each have advantages and limitations. The most traditional definition of a community could be described as a set of nodes that are more connected to each other than to the rest

of the network [26, 34, 35]. This is not the case for some more recent community detection frameworks, which are capable of finding groups of nodes which interact with each other and other communities in a statistically similar way but with no constraint related to a higher degree of connection within the community than the rest of the graph [21, 30, 36]. In particular, this is true of a lineage of methods that has developed out of the practice of using Stochastic Block Models in a generative framework to estimate modular structure. This was seemingly born out of the practice of generating benchmark networks to evaluate community detection performance, hence inference frameworks which use such models perform well when attempting to recover the communities under this definition. Although there are special cases where SBMs can be uses to find assortative communities explicitly [37, 38] these methods have not yet been implemented for weighted graphs. Hence, although the more principled SBM approach to community detection offers a stable and principled methodology for detecting communities in networks[36], the current accessible implementations are not appropriate for this application. Instead, I chose a modularity maximisation framework as the community definition is explicit in this method, which makes interpretation of the resulting communities more straightforward in this case as I explicitly sought assortative communities, which is problematic for, generally more robust, frameworks based on statistical inference[37, 38].

The chief limitation of the Leiden algorithm is a resolution limit which limits the algorithms ability to detect small communities [39], hence smaller communities within the partitions detected may exist. As such there could be other communities within provinces which are formed predominantly of particular religious groups. For the purpose of this analysis the scale of the communities observed was deemed appropriate and meaningful. The observation of strong connectivity between schools affiliated with the

Dutch reformed church (Reformatorisch) over multiple provinces is informative regardless of whether smaller community structures also exist.

To conclude, there are important correlations between religious faith and vaccination refusal in the Netherlands. In particular the traditions associated with low vaccine uptake are the Dutch Orthodox Reformed Church and the Anthroposophic community. The popularity of faith-based schools in the Netherlands can lead to schools with low overall vaccine uptake. Our network analysis of the connections between schools through shared households reveals that there are stronger connections between schools of the same denomination than would be expected at random. This is particularly clear within the Dutch Reformed Church. Although present, the effect is weaker for Anthroposophic schools.

The isolation of these particular socio-religious groups may have implications for the epidemiology of outbreaks in the Netherlands as a whole, particularly for measles, where vaccination uptake must be very high (~95%) to interrupt transmission effectively. Moreover, looking to the conclusions in Analysis A, the relative differences in vaccine uptake by religious affiliation and variation in isolation from the general population of schools may introduce substantial differences in risk to unvaccinated children depending on their religious group as a product of the overall network over and above the vaccine uptake in the particular school.

The nature of outbreaks on this network and the implications of the findings of this chapter are assessed in chapter 8 by means of outbreak simulation studies employing a disease transmission model constructed using this network.

7.5 References

1. Van Lier EA, Oomen PJ, Giesbers H, Conyn-van Spaendonck MAE, Drijfhout IH, Zonnenberg-Hoff IF, et al. **Vaccinatiegraad Rijksvaccinatieprogramma Nederland Verslagjaar 2014**. 2014. www.rivm.nl. Accessed 5 Dec 2019.

2. Velzen E van, Coster E de, Binnendijk R van, Hahné S. **Measles outbreak in an anthroposophic community in The Hague, The Netherlands, June-July 2008**. *Eurosurveillance*. 2008, 13:18945. doi:10.2807/ese.13.31.18945-en.

3. Woudenberg T, van Binnendijk RS, Sanders EAM, Wallinga J, de Melker HE, Ruijs WLM, et al. **Large measles epidemic in the Netherlands, May 2013 to March 2014: changing epidemiology.** *Euro Surveill*. 2017, 22. doi:10.2807/1560-7917.ES.2017.22.3.30443.

4. Nic Lochlainn LM, Woudenberg T, van Lier A, Zonnenberg I, Philippi M, de Melker HE, et al. **A novel measles outbreak control strategy in the Netherlands in 2013–2014 using a national electronic immunization register: A study of early MMR uptake and its determinants**. *Vaccine*. 2017, 35:5828–34. doi:10.1016/J.VACCINE.2017.09.018.

5. Ruijs WLML, Hautvast JLLA, Van Der Velden K, De Vos S, Knippenberg H, Hulscher MEEJL. **Religious subgroups influencing vaccination coverage in the Dutch Bible belt: an ecological study**. *BMC Public Health*. 2011, 11:102. doi:10.1186/1471-2458-11-102.

6. Fournet N, Mollema L, Ruijs WL, Harmsen IA, Keck F, Durand JY, et al. **Under-vaccinated groups in Europe and their beliefs, attitudes and reasons for non-vaccination; two systematic reviews**. *BMC Public Health*. 2018, 18:196. doi:10.1186/s12889-018-5103-8.

7. Ruijs WLM, Hautvast JLA, van IJzendoorn G, van Ansem WJC, van der Velden K, Hulscher ME. **How orthodox protestant parents decide on the vaccination of their children: a qualitative study**. *BMC Public Health*. 2012, 12:408. doi:10.1186/1471-2458-12-408.

8. van Lier A, van de Kassteele J, de Hoogh P, Drijfhout I, de Melker H. **Vaccine uptake determinants in The Netherlands**. *Eur J Public Health*. 2014, 24:304–9. doi:10.1093/eurpub/ckt042.

9. Klomp JHE, van Lier A, Ruijs WLM. **Vaccination coverage for measles, mumps and rubella in anthroposophical schools in Gelderland, The Netherlands**. *Eur J Public Health*. 2015, 25:501–5. doi:10.1093/eurpub/cku178.

10. Harmsen IA, Ruiter RAC, Paulussen TGW, Mollema L, Kok G, de Melker HE. **Factors that influence vaccination decision-making by parents who visit an anthroposophical child welfare center: a focus group study.** *Adv Prev Med*. 2012, 2012:175694. doi:10.1155/2012/175694.

11. Woudenberg T, Woonink F, Kerkhof J, Cox K, Ruijs WLM, Van Binnendijk R, et al. **Epidemiology and Infection The tip of the iceberg: incompleteness of measles reporting during a large outbreak in The Netherlands in 2013-2014**. *Epidemiol Infect*. 2018, 147:1–7. doi:10.1017/S0950268818002698.

12. European Centre for Disease Prevention. **Measles and Rubella Surveillance 2017**. doi:10.2900/11947.

13. Melegaro A, Jit M, Gay N, Zagheni E, Edmunds WJ. **What types of contacts are important for the spread of infections?: using contact survey data to explore European mixing patterns.** *Epidemics*. 2011, 3:143–51. doi:10.1016/j.epidem.2011.04.001.

14. Donker T, Wallinga J, Slack R, Grundmann H. **Hospital Networks and the Dispersal of Hospital-Acquired Pathogens by Patient Transfer**. 2012. doi:10.1371/journal.pone.0035002.

15. Donker T, Henderson KL, Hopkins KL, Dodgson AR, Thomas S, Crook DW, et al. **The relative importance of large problems far away versus small problems closer to home: insights into limiting the spread of antimicrobial resistance in England**. *BMC Med*. 2017, 15:86. doi:10.1186/s12916-017-0844-2.

16. Donker T, Smieszek T, Henderson KL, Johnson AP, Walker AS, Robotham J V. **Measuring distance through dense weighted networks: The case of hospital-associated pathogens**. doi:10.1371/journal.pcbi.1005622.

17. Donker T, Wallinga J, Grundmann H. **Patient Referral Patterns and the Spread of Hospital-Acquired Infections through National Health Care Networks**. *PLoS Comput Biol*. 2010, 6. doi:10.1371/journal.pcbi.1000715.

18. Wetenschap Ministerie van Onderwijs en Cultuur. **Leerplicht en kwalificatieplicht**. 2019. https://www.rijksoverheid.nl/onderwerpen/leerplicht/leerplicht-en-kwalificatieplicht. Accessed 5 Dec 2019.

19. Hagberg AA, Schult DA, Swart PJ. **Exploring network structure, dynamics, and function using NetworkX**. In: 7th Python in Science Conference (SciPy 2008). 2008. p. 11–5. http://conference.scipy.org/proceedings/SciPy2008/paper_2/. Accessed 6 Dec 2019.

20. Python Software Foundation. **Python Language Reference, version 2.7**. Python Software Foundation. 2013.

21. Fortunato S, Hric D. **Community detection in networks: A user guide**. *Phys Rep*. 2016, 659:1–44. doi:10.1016/j.physrep.2016.09.002.

22. Danon L, Díaz-Guilera A, Duch J, Arenas A. **Comparing community structure identification**. *J Stat Mech Theory Exp*. 2005, 2005:P09008–P09008. doi:10.1088/1742-5468/2005/09/P09008.

23. Traag VA, Waltman L, van Eck NJ. **From Louvain to Leiden: guaranteeing well-connected communities**. *Sci Rep*. 2019, 9:5233. doi:10.1038/s41598-019-41695-z.

24. Ronhovde RKDDRRP, Nussinov Z. **An edge density definition of overlapping and weighted graph communities**. 2013. http://arxiv.org/abs/1301.3120.

25. Li Z, Zhang S, Wang R-S, Zhang X-S, Chen L. **Quantitative function for community detection**. *Phys Rev E*. 2008, 77:036109. doi:10.1103/PhysRevE.77.036109.

26. Newman MEJ, Girvan M. **Finding and evaluating community structure in networks**. *Phys Rev E*. 2004, 69:026113. doi:10.1103/PhysRevE.69.026113.

27. Aldecoa R, Marín I. **Surprise maximization reveals the community structure of complex networks**. *Sci Rep*. 2013, 3:1060. doi:10.1038/srep01060.

28. Peixoto TP. **Revealing consensus and dissensus between network partitions**. 2020. http://arxiv.org/abs/2005.13977. Accessed 6 Oct 2020.

29. Lancichinetti A, Fortunato S. **Consensus clustering in complex networks**. *Sci Rep*. 2012, 2:336.

doi:10.1038/srep00336.

30. Yang Z, Algesheimer R, Tessone CJ. **A Comparative Analysis of Community Detection Algorithms on Artificial Networks**. *Sci Rep*. 2016, 6:30750. doi:10.1038/srep30750.

31. Coleman J. **Relational Analysis: The Study of Social Organizations with Survey Methods**. *Hum Organ*. 1958, 17:28–36.

32. Hahne S, te Wierik MJM, Mollema L, van Velzen E, de Coster E, Swaan C, et al. **Measles Outbreak, the Netherlands, 2008**. *Emerg Infect Dis*. 2010, 16:567–9. doi:10.3201/eid1603.090114.

33. Peri-Rotem N. **Religion and Fertility in Western Europe: Trends Across Cohorts in Britain, France and the Netherlands**. *Eur J Popul*. 2016, 32:231–65. doi:10.1007/s10680-015-9371-z.

34. Danon L, Read JM, House TA, Vernon MC, Keeling MJ. **Social encounter networks: characterizing Great Britain.** *Proc Biol Sci*. 2013, 280:20131037. doi:10.1098/rspb.2013.1037.

35. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. **Fast unfolding of communities in large networks**. *J Stat Mech Theory Exp*. 2008, 2008:P10008. doi:10.1088/1742-5468/2008/10/P10008.

36. Abbe E. **Community Detection and Stochastic Block Models: Recent Developments**. *J Mach Learn Res*. 2018, 18:1–86. http://jmlr.org/papers/v18/16-480.html.

37. Zhang L, Peixoto TP. **Statistical inference of assortative community structures**. 2020. http://arxiv.org/abs/2006.14493. Accessed 6 Oct 2020.

38. Lu X, Szymanski BK. **A Regularized Stochastic Block Model for the robust community detection in complex networks**. *Sci Rep*. 2019, 9:1–9. doi:10.1038/s41598-019-49580-5.

39. Fortunato S, Barthelemy M. **Resolution limit in community detection**. *Proc Natl Acad Sci*. 2007, 104:36–41. doi:10.1073/pnas.0605965104.

8 Analysis D (part 2): A network of schools in the Netherlands: Implications for measles epidemiology

Objective: *Analyse the impact of faith schools on clustering of children who are susceptible to measles and resultant measles epidemiology in the Netherlands*

8.1 Introduction

Since the introduction of mumps, measles and rubella vaccine (MMR) in the Netherlands in 1989, sporadic measles outbreaks have persisted [1–5]. Most notably, in 1999 [6] and 2013 large outbreaks were recorded, both with c. 3000 cases reported and an estimated total incidence of around 30,000 cases [7]. The majority of reported cases (94%) were unvaccinated individuals, 84% were individuals who refused vaccination for religious or political reasons and 90% were children (< 20 years)[5].

Uptake of MMR in the Netherlands is generally high, with 95% uptake for the first dose of MMR by the age of 14 months and 93% uptake for the second dose of MMR by the age of 10 years [5, 8]. It is broadly understood that low vaccination coverage in particular socio-religious groups contributes substantially to the large outbreaks observed [9–15], however these groups are relatively sparsely distributed geographically. An outbreak of measles in such a highly vaccinated population would not be expected to infect such a large proportion of susceptible individuals without a high degree of clustering of unvaccinated people. Although the association of particular religious groups provides a basis for this clustering, the mechanism and extent of clustering has not been well enough quantified to explain the size and frequency of observed outbreaks. Previous modelling analyses [16] of these outbreaks have relied on strong assumptions of homogenous contact exclusively within unvaccinated communities to reproduce outbreak sizes observed. Moreover, although large outbreaks have been observed with an a period of 12 to 13 years, other smaller outbreaks have been recorded in the intermittent years[1–3]. It is unclear whether timing of the outbreak or location or faith affiliation of the initial cases has a part to play in the final size.

In Analysis D (part 1), I analysed the household based social contacts between schools in the Netherlands. I showed higher homophily and general contact network proximity between faith schools of the same denomination than between schools that did not share the same faith denomination. Schools that are affiliated with the Dutch Reformed Church, a denomination of the Christian faith who generally refuse vaccination, are particularly well connected to each other on the contact network. The results of Analysis D (part 1) indicate that schools and households may provide enough contact to connect a large proportion of unvaccinated children, particularly those in the Dutch Reformed community. But the questions remain:

Is the concentration of unvaccinated children in particular schools sufficient to explain the large outbreaks of measles observed in 1999/2000 and 2013/14? and how important is the clustering of unvaccinated children in particular schools and the network proximity of schools with low vaccination?

In this chapter I use the network of contact between schools to construct a model of measles transmission across the Dutch school system to address these questions.

8.2 **Methods**

I used the network described in Analysis D (part 1) to construct a model of transmission of infectious disease. By incorporating estimates of school level uptake, estimated by Klinkenberg et al [17], I simulated outbreaks of measles to assess the size and geographical characteristics of an outbreak that would be expected if contact were made

only through school and households between school aged children. I performed two simulation studies to:

1. Quantify the risk to (and posed by) each school in the network. Using this I evaluate how epidemics vary depending on the denomination of the school where they are initiated.

2. Evaluate how clustering of unvaccinated children in schools affiliated with particular faiths may have contributed to large outbreaks of measles virus in the Netherlands;

To evaluate the importance of clustering of unvaccinated children in particular schools (individual level clustering) and clustering in the network of schools with low vaccination (school level clustering) to the result of a simulated outbreak, I also performed the analysis in simulation studies A and B with two alternative networks. Firstly, with vaccination uptake in each school made equal to the uptake of the four-digit postcode areas (PC4s) of the children who attend that school. Secondly, I constructed a network with contact between the schools defined by a spatial kernel, as opposed to the network constructed from the national school data (figure 8.1).

Overview of school contact network

I used school and pupil data provided by the Dutch Ministry for Education (DUO), which holds data for each school (n. 9200) and individual child in the educational system on the 31st October 2013.

Using data on the residence and school of each pupil in the database I constructed a network [18] of schools where edges were weighted by the number of unique contact pairs, where a contact pair comprises two children who reside in the same household but attend different schools. This is discussed in detail in Chapter 5.

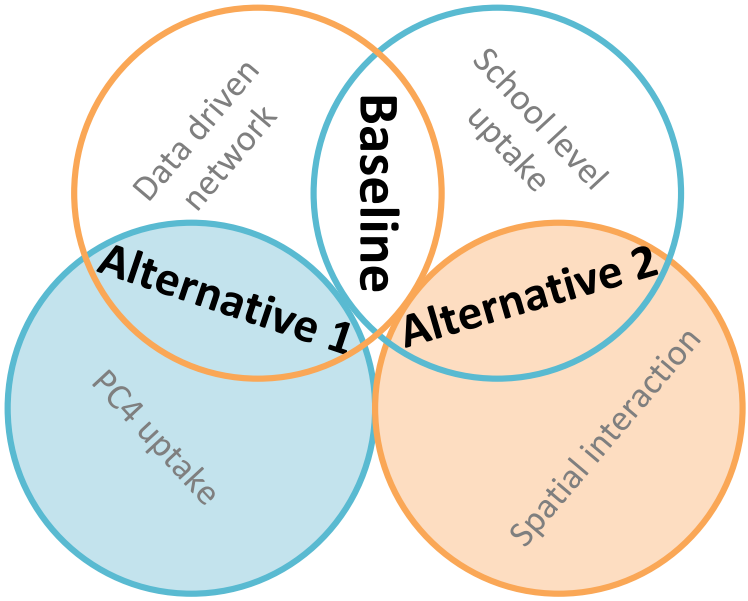


Figure 8.1 Schematic of the components of the different network models

The baseline used the data driven contact network and school level uptake. Alternative Models: 1. Uptake based on four-digit postcode areas of the children who attend the school. 2. Interaction based on a spatial interaction kernel.

Transmission model

Using the information in the school contact network, I constructed a transmission probability network, where edges were weighted with the probability that an outbreak would be initiated in school *i* in the event of an outbreak in school *j*.

I estimated probability of transmission between pairs of schools as follows:

First, I defined the probability of transmission between a single contact pair that form a link between school i and school j. Considering one contact pair: I set the probability, q, of transmission between the pair, given that the child in the pair who attends school j is infected and the other child, who attends school i, is susceptible. I denote the probability that the child from school j is infected to be P_j^I and the probability that the child from school i is susceptible to be P_i^S. The overall probability of transmission between a contact pair is given by the intersection of these:

$$P_i^I P_j^S q$$

The probability that the newly infected child causes an outbreak in school i is denoted by P_i^{OB}. The total probability of an outbreak being seeded in school i from an ongoing outbreak in school j over all C_{ij} contact pairs is given by:

$$P_{trans,ij} = 1 - \left(1 - P_j^I P_i^S q P_i^{OB}\right)^{C_{ij}}$$

I estimated the probability that the child in school j is infected by the outbreak, P_j^I as the proportion of the school children infected by the outbreak ($R_j(\infty)$) in that school, which I calculated by solving the final size equation [19]:

$$R_j(\infty) = \left(1 - V_j\right)\left(1 - e^{-\left(1-V_j\right)R_0 R(\infty)}\right)$$

where V_j is the vaccination coverage in school j. I estimated the probability that the child in school i is susceptible, P_i^S, to be equal to the proportion of school i that are unvaccinated according to the inferred school level uptake rates, $(1 - V_i)$.

I took the probability of an outbreak in that school as a result of a single transmission event to be:

$$P_i^{OB} = \left(1 - \frac{1}{R_e}\right) = \left(1 - \frac{1}{(1 - V_i)R_0}\right)$$

which assumes a geometric distributed contact rate and homogenous mixing amongst children within schools[19]. Note that a key difference between this model and the one used in Analysis C is the explicit presence of vaccination coverage. Because of different vaccination coverage in different schools, the transmission probability network is now a "directed graph", i.e. $P_{trans,ij} \neq P_{trans,ji}$.

School level vaccine uptake estimates

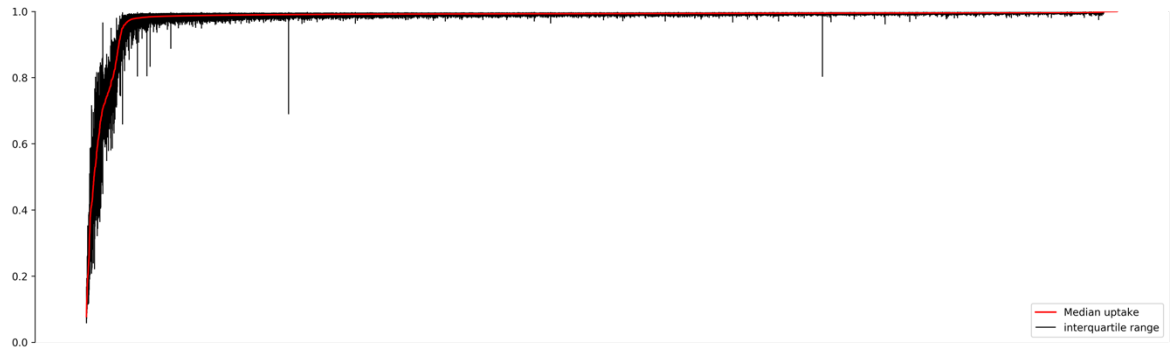
Estimates of vaccine uptake in schools were taken from a separate analysis by Klinkenberg et. al. [17]. In this analysis, which is yet to be fully published, school level uptake was estimated using a hierarchical Bayesian framework incorporating vaccine uptake data at postcode level collected in 2014 [8], school catchment data from 2013 to 2016 (at postcode level) and data on the transition of pupils from primary to secondary schools for the same years. The authors evaluated their estimates against measured uptake in schools in Utrecht, which demonstrated good agreement with the model. The results of this analysis quantified and corroborated other evidence of low uptake of MMR

vaccine in schools associated with Dutch Reformed and Anthroposophic socio-religious groups (figure 8.2). With agreement from the authors, I used the joint posterior distribution of the vaccine uptake by school from their analysis to parameterise the simulations.

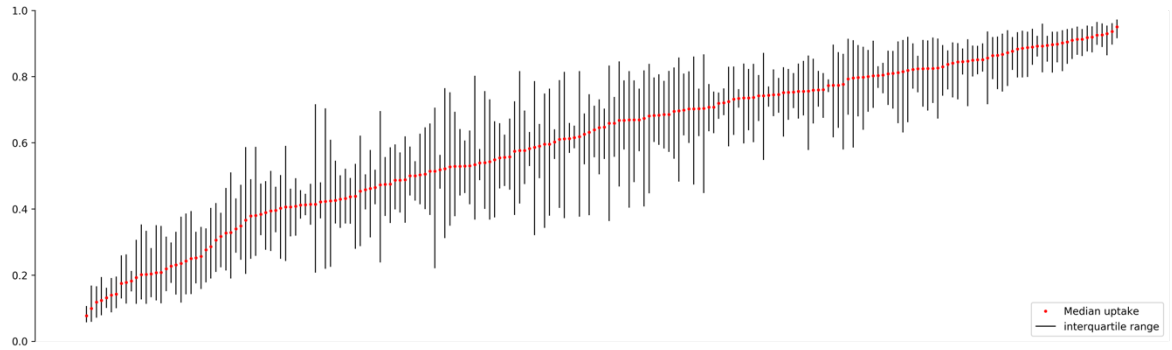
Alternative Model 1: Approximating school level vaccination from areal vaccination data.

In the baseline model unvaccinated children were highly concentrated in particular schools, as quantified by Klinkenberg et al [17]. To test the importance of high clustering of unvaccinated children in particular schools to outbreak size and distribution, I analysed an alternative parameterization of the model where vaccination in each school was estimated using the vaccination rates of the PC4s (four-digit post code areas) of the children in attendance at the school. Assuming children had a probability of being vaccinated equal to the vaccine uptake of the PC4 where they lived. I used data provided by DUO on the residence of children in each school to calculate the proportion of children in each school who live in each PC4. School vaccination uptake was set as the weighted average of vaccination uptake rates at PC4 level, weighted by the proportion of children who live in each PC4.

A



B



C

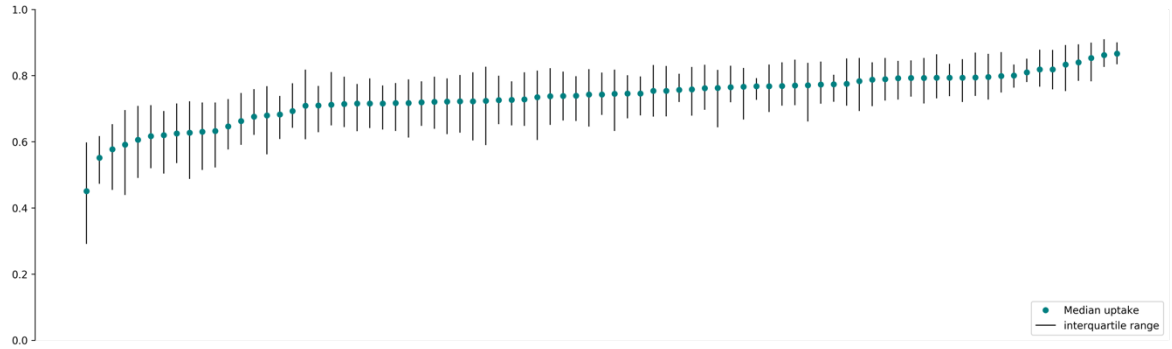


Figure 8.2 Ranked vaccine uptake in schools, points show mean, bars show interquartile range of the marginal distribution for each school.

A: all schools, B: Dutch Reformed schools, C: Anthroposophic schools.

Alternative Model 2: Spatial interaction network

Analysis D (part 1) shows that due to homophily between schools of particular denominations, the specific local structure of the network increases clustering of schools that are likely to have low vaccine uptake. To assess the importance of the specific network structure defined by the school data to the overall dynamics of outbreaks, I constructed an alternative school contact network where the geographic distance between connected schools followed a similar relationship to the baseline model, however contact is spread evenly over all schools according to that relationship.

The spatial distribution of a school's 'neighbours' can reasonably be described by a semi-Gaussian spatial relationship. I used this relationship weighted by the degree of the connecting schools. To calibrate the spatial kernel, I matched the distribution of distance between schools connected by contact pairs to the school data derived network Appendix E. Interaction between all schools was calculated equally, regardless of primary and secondary status (Figure 8.3).

Evaluation of school level infectivity

To evaluate the infectivity of each school in the network, I calculated the weighted out-degree of each school for multiple realisations of the transmission probability network. This value is the sum of the out-edges of each school and amounts to the expected number of adjacent schools infected if an outbreak were initiated in that school.

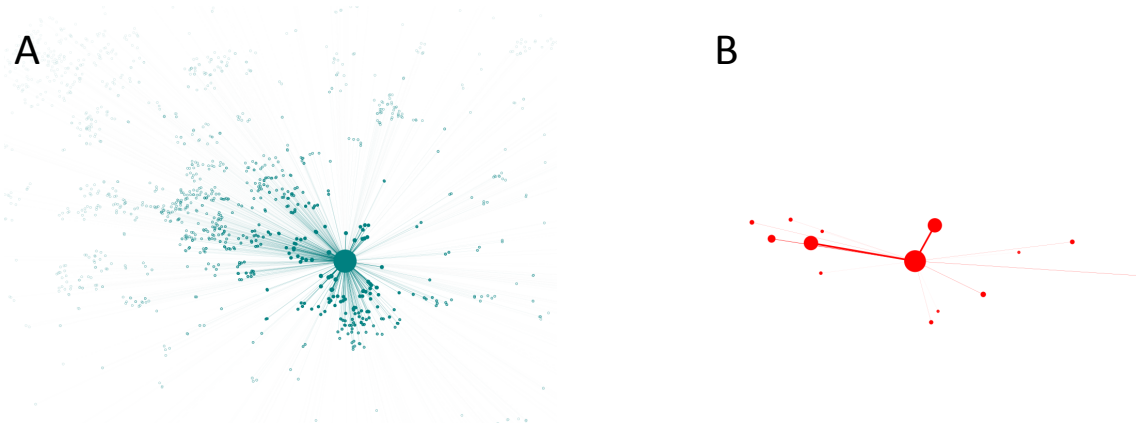


Figure 8.3 Ego-networks of the school where the 2013/14 measles outbreak was seeded

Position of the nodes shows the relative location of the schools, node size gives their weighted degree (the number of unique contact pairs with the seed school), the edge width indicates the number of unique contact pairs between the schools. A. Connections are based on spatial interaction with all schools. B. Connections are based on the school data.

Outbreak simulations

I made the assumption that schools were one of: susceptible, infected or recovered. Each susceptible school, i, has susceptibility equal to $1 - V_i$. Infected schools are those affected by an outbreak, and have a probability of infecting neighbouring susceptible schools ($P_{trans,ij}$) as defined above. After an outbreak has occurred, I assumed that the school had effectively depleted its susceptible population, and could not be re-infected.

Once vaccine uptake is assigned to each school (e.g. sampled from the joint posterior distribution provided by Klinkenberg et.al.), probability of transmission between schools is constant resulting in a static directed network, where out-edges (probability of outward transmission from a school to its neighbour) are not necessarily equal to the equivalent in-edge (probability of inward transmission from a neighbour to a school).

For each set of school uptake values, I created 1000 instances of a *directed binary outbreak network* from the transmission probability network. Where a *directed binary outbreak network* is a network with edges equal to 1 or 0 between schools, where an edge of 1 from school i to school j indicates transmission would occur from school i to school j in the event of an outbreak in school i. For each instance, each edge is weighted 1 with probability informed by the equivalent edge value in the transmission probability network.

Note, these networks differ from those discussed in Analysis C only because of the different vaccine uptake in each school. This difference in susceptible fraction necessitates the directed nature of the *transmission probability network* and the *binary outbreak network*. In Analysis C these were undirected as total susceptibility was assumed.

Directed binary outbreak networks provide the information required to easily identify both:

A) *Schools at risk of infection from each school*: The set schools which would be infected if an outbreak initiated in each particular school in the network (i.e. all the schools that would be infected by an outbreak initiated in school *i*). For each particular initial school, I identified schools that would be infected by an outbreak as those connected by chains of out-edges (out-component). i.e. each generation of the outbreak is comprised of the successors (figure 8.4) of the schools in the previous generation. (Appendix E)

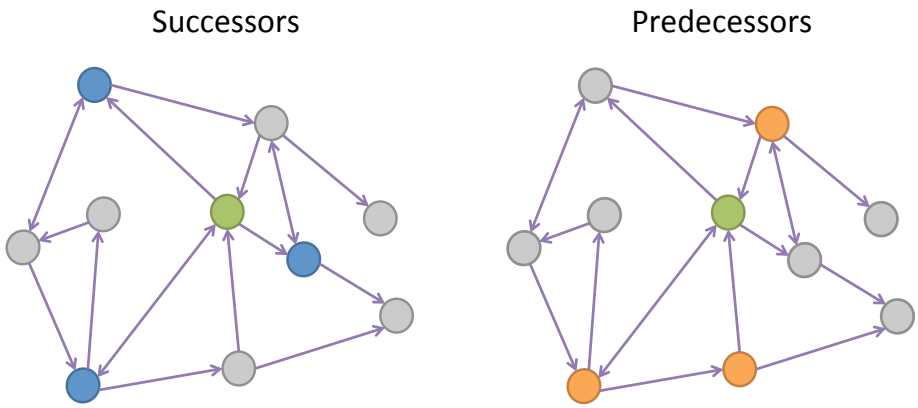


Figure 8.4 Successors and predecessors in a directed network.

A: The successors of a particular node, indicated by green, would be infected if an outbreak occurred in it. B: Predecessors, orange nodes, are those schools that would infect a particular school, indicated by green node, if there were an ongoing outbreak within their population.

B) *Schools from which each school is at risk of infection*: The set of schools within which if an outbreak were initiated, it would lead to infection in each particular school (i.e. all the schools that would lead to school *i* becoming infected if an outbreak initiated in them). I identified these schools as those connected to the school by chains of in-edges (in-component). That is, schools that would infect each school in-edge of value 1 connecting them, i.e. the predecessors of that school. For each school I identified the schools whose outbreak would eventually infect that school. (Appendix E)

Calculating average number of cases per postcode area (PC4)

Data was made available to us by DUO detailing the number of children in each school whose residential address falls within each PC4. For each instance of the model I calculated the expected number of cases per PC4 (FS_{pc4}) as the sum of the proportion of students in each infected school who reside in that PC4, multiplied by the proportion of the school that was infected in the outbreak.

$$FS_{pc4=p} = \sum_i P_{p,i} R_i(\infty)$$

Where, $P_{p,i}$ is the proportion of school i residing in PC4 p, and $R_i(\infty)$ is the proportion of students in school i who were infected in the outbreak.

Simulation study 1: Risk by School

To quantify the risk posed by each school in the network, I calculated the mean expected final outbreak size (schools and children) in the event of an outbreak initiated in each school in the network. To quantify the risk posed to each school in the network I calculated the number of schools in which an outbreak could be initiated in that could lead to infection in the school. I weighted each school's contribution to risk by the probability of an outbreak successfully taking place. I took this to be the proportion of all unvaccinated children in the network who attend that school in the school, multiplied by the probability of a child seeding an outbreak $1 - 1/R_{eff}$.

I compared the risk posed by and to Dutch Reformed and Anthroposophic schools by evaluating risk relative to vaccine uptake in these denominations. To quantify the importance of clustering of schools of a particular denomination in the network, I repeated this analysis for both the baseline model and Alternative Model 2, which uses spatial interaction between schools instead of household links.

Simulation study 2: Simulating the 2013/14 Outbreak: Evaluating the Model

To evaluate the ability of the school network transmission model to describe observed outbreak dynamics, I analysed 1000 simulated outbreaks initiated in the schools where the first cases were reported in the 2013 outbreak.

To compare the geographic distribution of cases in 2013/14 to those predicted by each model, I used a Receiver Operating Characteristic (ROC) at PC4 level. Where true positive PC4s are those where cases were predicted by the model, and also cases were observed in 2013 etc.

Sensitivity is calculated as the proportion of PC4s that had cases reported, which the model also predicted cases in. Specificity is calculated as the proportion of PC4 areas where cases were predicted, that also had cases reported in the outbreak.

To reflect the relative importance of PC4s with higher reported or simulated incidence, I also calculated a weighted ROC (wROC). For this measure, when calculating sensitivity, each PC4 with cases reported is weighted by its proportional contribution to all cases reported. Hence, whereas for the unweighted ROC the sensitivity value is the proportion of areas with cases reported that also had cases predicted, for the wROC the sensitivity value is the proportion of cases reported that occurred within PC4s where cases were predicted by the model.

In addition, when calculating specificity, each PC4 area with cases predicted, is weighted by the proportion of cases predicted in the total model output located in that PC4 area.

Hence, the weighted specificity gives the proportion of cases predicted that were in PC4 areas where cases were reported in the outbreak.

To quantify the importance of clustering of unvaccinated children in school and clustering of schools in the broader network for explaining the size and geographical distribution of cases in 2013/14, I evaluated the performance of the baseline model (school data derived network with vaccination uptake informed by Klinkenberg et al) relative to the performance of Alternative Model 1 (PC4 level vaccination uptake) and Alternative Model 2 (Spatial interaction between schools).

8.3 Results

Simulation study 1: Risk posed by school - Final size by initial school

The overall risk posed by an outbreak in each particular school was quantified by finding the distribution of final outbreak size. For both the school data and spatial networks, the majority of schools had a very low mean outbreak size, as no sustainable transmission was observed in any simulation.

For the baseline model, the maximum mean outbreak size was 171 schools and 23,766 children and was associated with a Dutch Reformed school (Figure 8.5). In general, Dutch Reformed schools generated large outbreaks particularly for schools with very low vaccination coverage. Outbreaks seeded in Anthroposophic schools generally remained much smaller, with a maximum of 5 schools and 616 children.

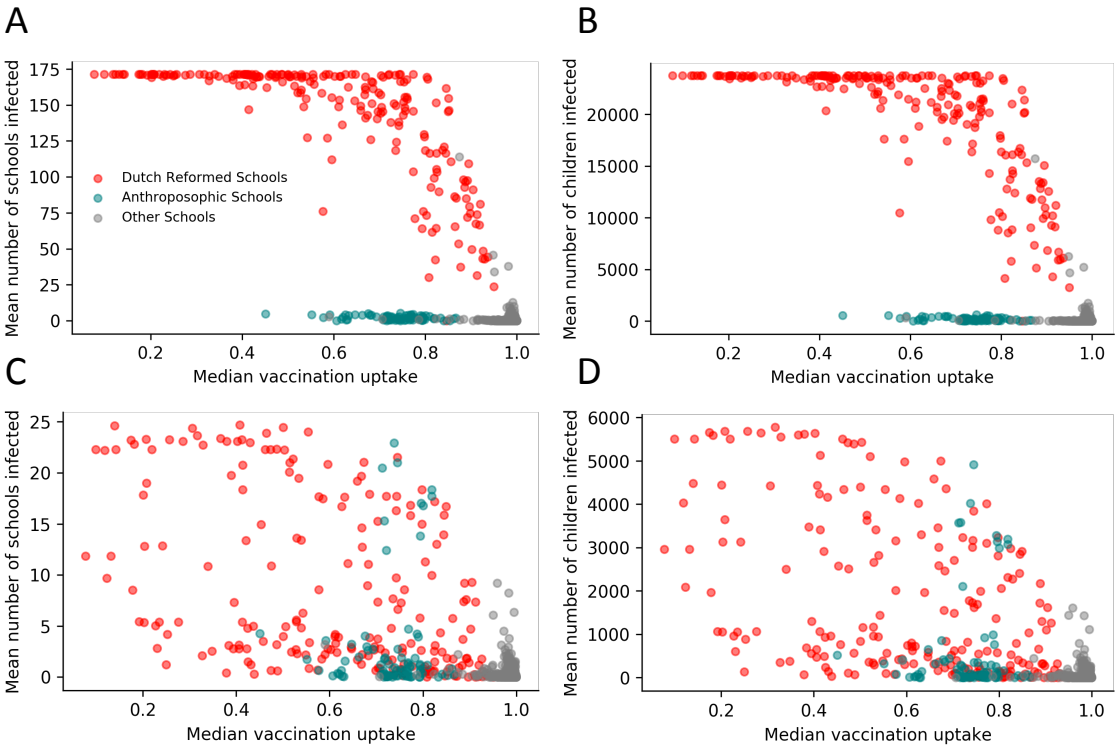


Figure 8.5 Mean outbreak final size by school where outbreak is seeded.

Red points indicate Dutch Reformed schools, green points indicate Anthroposophic schools, grey points indicate other schools. A: School data derived network- mean number of schools infected, B: School data derived network- mean number of children infected, C: Spatially derived model- mean number of schools infected, D: Spatially derived model- mean number of children infected.

Outbreaks simulated on Alternative Model 2 were generally much smaller. There was a trend with vaccine uptake, however there remained schools where vaccine uptake was low that still only seed very small outbreaks. The schools that seeded the largest outbreaks using Alternative Model 2 had a mean of 25 infected schools and 5782 cases. The Dutch reformed schools seeded the largest outbreaks. However, the difference between Dutch Reformed and Anthroposophic schools was much less substantial with a number of Anthroposophic schools seeding outbreaks greater than predicted by the model derived from the school data, with a maximum of 23 schools and 4916 children. Notably the some

Anthroposophic schools seeded outbreaks comparable to those seeded by Dutch Reformed schools with similar vaccine uptake.

The risk posed to particular schools is summarised as the mean proportion of possible initial cases in unvaccinated children that would eventually lead to an outbreak in the school (i.e. the number of children who could be infected and initiate an outbreak which reaches the school). This shows a close relationship to the risks posed by the school in terms of the expected outbreak sizes. The highest risk school would have an outbreak seeded as a result of 40% of seed cases (Figure 8.6); this was a Dutch Reformed school. The mean proportion of seeds that would cause an outbreak in Dutch reformed schools was 33%. The risk posed to Anthroposophic schools was much lower, with a maximum and mean proportion of seeds leading to outbreaks in particular schools at 1% and 0.3% respectively.

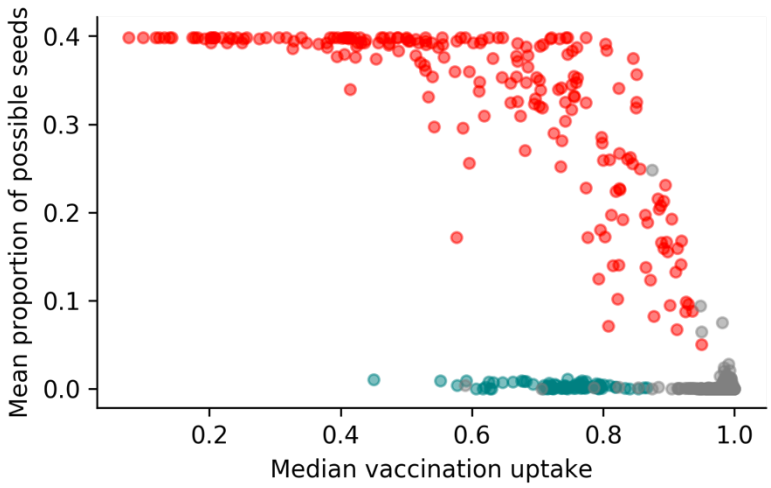


Figure 8.6 The proportion of unvaccinated children who if seeded an outbreak in their school, would cause an outbreak in each school plotted.

Red indicates Dutch Reformed schools, green indicates Anthroposophic schools and grey indicates other schools.

Simulation study 2: Recreating the 2013/14 outbreak

General description of the outbreaks predicted by each model

Using the baseline model (National school data contact network and school level vaccine uptake estimates), 1000 outbreak simulations with the initial schools set to the two schools where the first cases were reported in the 2013 outbreak, resulted in a mean of 28576 (28022 – 29078 95% CI) cases. The geographical distribution of cases was broadly consistent with the reported cases in 2013/14. There was a high likelihood of cases being reported in PC4 areas in the centre of the country and the southwest. There was also a high likelihood of detecting cases in a small region in the north east of the country (figure 8.7 C).

When Alternative Model 1 was used, (national school data in combination with the vaccination uptake in schools estimated from PC4 level vaccine uptake), the mean final size of the outbreaks was 9093 (420 – 18209 95% CI). The cases were distributed in a narrow strip, with high frequency of cases stretching from the south west region to the north east of the central region (figure 8.7 B).

When Alternative Model 2 was used (the spatially derived contact network and school level vaccine uptake estimates), the final size of the outbreak was 67 (10 – 163 95% CI) cases. The majority of cases predicted occurred in schools in the central region of the country, with low probability of detecting infection in any other regions (figure 8.7 A).

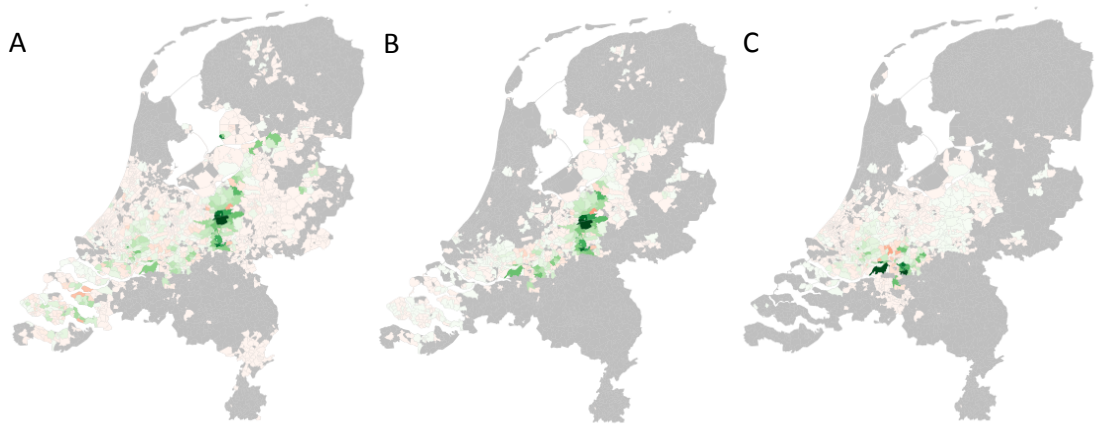


Figure 8.7 Mean number of cases across 1000 simulated in each PC4 region with a reporting rate of 10% (from estimates in literature).

A: the baseline model: School data network with school level uptake., B: Alternative Model 1: School data network with PC4 level uptake, C: Alternative Model 2: Spatial network with school-level uptake

Model assessment using Receiver Operating Characteristic (ROC) and weighted ROC

Using the unweighted ROC, the mean sensitivity (proportion of *PC4s* where cases reported that were predicted by the model) was 0.84, 0.28 and 0.18, for the Baseline model, Alternative Model 1 and Alternative Model 2 respectively (Figure 8.8 A). The mean specificity (proportion of *PC4s* where cases were predicted that also had cases reported) was 0.40, 0.54 and 0.57 for Baseline model, Alternative Model 1 and Alternative Model 2 respectively.

Considering the weighted ROC, the mean sensitivity (proportion of *cases* reported that were in PC4 areas predicted by the model) was 0.94, 0.38 and 0.23 for the Baseline model, Alternative Model 1 and Alternative Model 2 respectively (Figure 8.5 B). The mean specificity (proportion of *cases* predicted that fell into PC4 regions where cases were reported) was 0.94, 0.91 and 0.91 for School data network with school vaccination,

Spatial network with school vaccination and School data with PC4 level vaccination respectively.

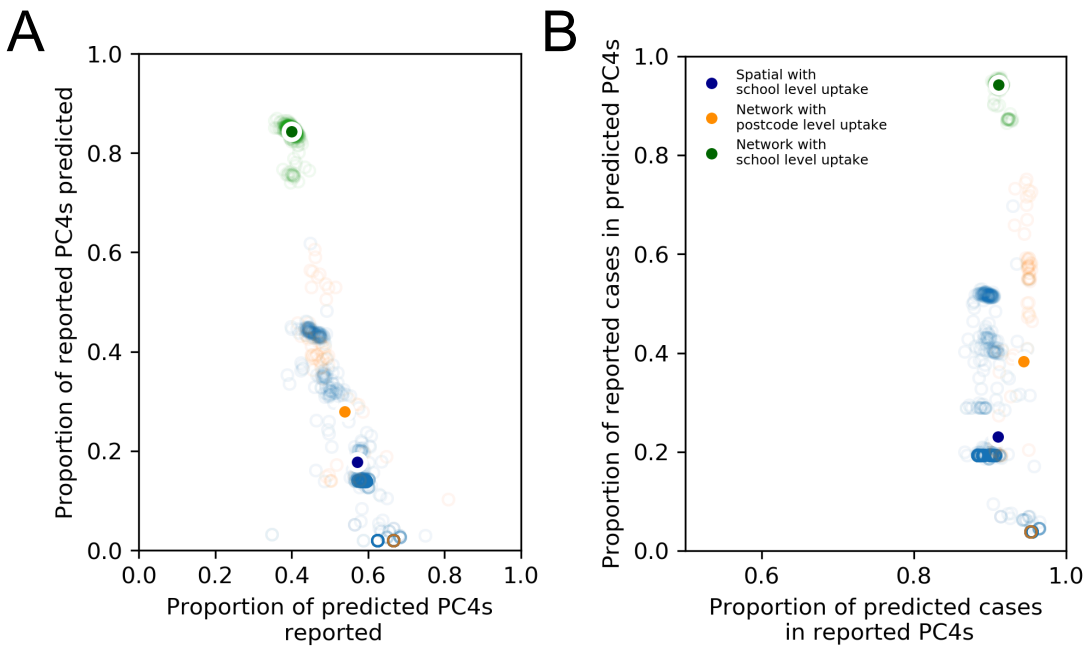


Figure 8.8 Sensitivity and specificity of the baseline and alternative network models.

A: Unweighted sensitivity vs. specificity. B: Weighted sensitivity vs. specificity.

8.4 **Discussion**

Despite high-average uptake of MMR vaccination over the past two decades. the Netherlands continues to experience large outbreaks of measles[1–5]. In general, these outbreaks are associated with socio-religious groups who refuse to vaccinate their children in large numbers. However, the geographical scale and total incidence of the outbreaks varies considerably between socio-religious groups, with outbreaks in the Dutch Reformed population generally being much larger and farther reaching than those in the Anthroposophic population. I used the Dutch national network of schools

constructed from national school data, as a framework for simulating outbreaks of measles, seeded in Dutch Reformed, Anthroposophic and other schools in the network.

Over two simulation studies of the network my model suggests that the school network has the capacity to explain key differences between outbreaks in Dutch Reformed and Anthroposophic populations, and can offer important insight into the epidemiology of measles in the Netherlands over the past 20 years, and the potential for large outbreaks to occur in the future.

I have made observations, which have important implications for understanding the determinants of large outbreaks of measles in the Netherlands and their timing.

Firstly, the results from spatially derived model suggests that high rates of community transmission and long-range transmission would be required to recreate the observed epidemiology as schools with low vaccination uptake were less clustered within the network. Although the data derived network and postcode level vaccination better explain the geographical distribution of cases than the spatially derived network, it falls short of the model using the same contact network with school level vaccination. In addition, the overall size of the outbreak is not reflected, suggesting that clustering of unvaccinated children in schools is important for explaining the full extent of the outbreak observed. These findings indicate that the distribution of unvaccinated children within particular schools, and the specific links between these schools greatly increase the potential for large outbreaks to occur. With these factors accounted for in the model, outbreaks similar to that observed in 2013/14 can be simulated by accounting for school and household transmission only. This finding suggests that, in a population with immunity provided

only by vaccination, outbreaks have a clear determinable reach, which is largely unaffected by chance encounters or rare long-range transmission events.

Secondly, the specific connections between schools described in the network were able to explain the difference in outbreak size in the different socio-religious groups. Outbreaks seeded in Dutch Reformed schools frequently resulted in large outbreaks infecting over 100 schools. Using the data derived network, the outbreaks seeded in Anthroposophic schools remained small even when the vaccine uptake in the seed school was low. In the spatial comparison model, the outbreaks seeded in the Dutch Reformed schools were much smaller and more comparable to those seeded in Anthroposophic schools. These findings suggest that the structure of the school system provides a mechanistic explanation for the difference in outbreaks observed in the past decade. Further, since the variation in outbreak size are due to structural differences in the population, it is likely that future outbreaks in these communities would follow similar patterns, if the structure of the school system remains comparable in years to come.

The purpose of this analysis was to identify whether the observed epidemiology of measles in the Netherlands is reproducible under these assumptions when explicit links between schools and households are accounted for in a model. This approach makes some important simplifying assumptions that the majority of transmission of measles between children occurs between contacts that either reside in the same home or attend the same school.

First, the model made some assumptions about transmission routes in the population. The model does not model transmission between school-aged children that do not have contact

in school or at home. In addition, the model does not account for transmission outside of the school-aged population. In reality, adults and preschool-age infants are likely to contribute to transmission to some degree. These neglected routes of transmission could potentially influence transmission dynamics in a way that this model cannot capture. In 2013/14 there were 438 cases (19%) in children between 1 and 4, lower than the 819 (30%) and 868 (32%) cases in 5-9 and 10-14-year-old age groups, suggesting less but by no means negligible transmission within pre-school-age than school-aged children[5]. The presence of pre-school institutions in the network would, most likely, provide additional connectivity on the network. This may increase transmission opportunities between primary schools in particular.

Secondly, my model does not simulate within-school transmission dynamics, but instead assumes a deterministic final size, which occurs with a probability determined by the effective reproduction number in that school. This cannot capture the contribution of outbreaks that do not reach sustainable transmission within schools, but still represent some small risk in terms of infecting other schools with the few pupils that are infected. Furthermore, the final size approximation I used assumes frequency dependent, homogenous transmission within the school population, which is a simplification. As discussed in previous chapters of this thesis, several surveys of school contacts reveal strong preference for mixing within school years [20–22]. This is likely to impact final size somewhat, although the impact of this on the propensity for outbreaks to pass between schools is likely to be small; due to the high infectiousness of measles, very few susceptible-infectious pairs between schools are required to successfully seed a new outbreak.

Thirdly, similarly to Chapter 6 of this thesis, the framework uses the same value of R_0 in schools and transmission probability q in all households. This is unrealistic, however any impact of variation in R_0 is likely to be negligible compared to the variation in vaccine uptake observed in this setting. Furthermore, at high values of R_0 and q that are consistent with measles transmission, small variations have little impact on the transmission probability calculated between schools.

Finally, the model works purely on a 'generational' basis, with no explicit temporal element. This restricts its use to modelling the overall incidence of an outbreak without modelling the temporal dynamics. This precluded the option of including weekends and school holidays in the model, where contact patterns have been shown to differ substantially [23–25]. Seasonal events have also been shown to affect the spatial distribution of cases, which has not been captured here [26].

Although these limitations are important for interpretation of the model output, they do not detract from the findings that the school system provides a system of contact that is able to facilitate large outbreaks amongst unvaccinated children in the Dutch Reformed population but not amongst children in Anthroposophic schools.

Further analysis of this network could allow study of other infectious diseases such as mumps and rubella, which are also prevalent amongst school-aged children within the same socio-religious populations. The model could also be extended to analyse outbreaks of influenza, where a large degree of transmission occurs within school age children. Another use of this framework could be to evaluate the effectiveness of various other intervention strategies, such as school closure. This method could also be applied in other

settings where vaccine uptake is strongly related to particular social groups[11], although this relies on detailed schools data being made available.

In conclusion, explicitly modelling connections between schools can provide important insights into the epidemiology of measles in the Netherlands, why it may vary between socio-religious groups. The results suggest that the preference for particular faith schools amongst families that choose not to vaccinate provides a mechanism for the clustering of unvaccinated children within the Dutch Reformed schools across the Netherlands. The same effect is not present in children that belong to the Anthroposophic community, resulting in much smaller outbreaks in this population.

8.5 References

1. Wallinga J, Heijne JCM, Kretzschmar M. **A measles epidemic threshold in a highly vaccinated population.** *PLoS Med*. 2005, 2:e316. doi:10.1371/journal.pmed.0020316.

2. Velzen E van, Coster E de, Binnendijk R van, Hahné S. **Measles outbreak in an anthroposophic community in The Hague, The Netherlands, June-July 2008**. *Eurosurveillance*. 2008, 13:18945. doi:10.2807/ese.13.31.18945-en.

3. Hahne S, te Wierik MJM, Mollema L, van Velzen E, de Coster E, Swaan C, et al. **Measles Outbreak, The Netherlands, 2008**. *Emerg Infect Dis*. 2010, 16:567–9. doi:10.3201/eid1603.090114.

4. van den Hof S, Meffre CM, Conyn-van Spaendonck MA, Woonink F, de Melker HE, van Binnendijk RS. **Measles outbreak in a community with very low vaccine coverage, The Netherlands.** *Emerg Infect Dis*. 2001, 7 3 Suppl:593–7. doi:10.3201/eid0707.010743.

5. Woudenberg T, van Binnendijk RS, Sanders EAM, Wallinga J, de Melker HE, Ruijs WLM, et al. **Large measles epidemic in The Netherlands, May 2013 to March 2014: changing epidemiology.** *Euro Surveill*. 2017, 22. doi:10.2807/1560-7917.ES.2017.22.3.30443.

6. van den Hof S, Meffre CM, Conyn-van Spaendonck MA, Woonink F, de Melker HE, van Binnendijk RS. **Measles outbreak in a community with very low vaccine coverage, The Netherlands.** *Emerg Infect Dis*. 2001, 7 3 Suppl:593–7. doi:10.3201/eid0707.010743.

7. Woudenberg T, Woonink F, Kerkhof J, Cox K, Ruijs WLM, Van Binnendijk R, et al. **Epidemiology and Infection The tip of the iceberg: incompleteness of measles reporting during a large outbreak in The Netherlands in 2013-2014**. *Epidemiol Infect*. 2018, 147:1–7. doi:10.1017/S0950268818002698.

8. Van Lier EA, Oomen PJ, Giesbers H, Conyn-van Spaendonck MAE, Drijfhout IH, Zonnenberg-Hoff IF, et al. **Vaccinatiegraad Rijksvaccinatieprogramma Nederland Verslagjaar 2014**. 2014. www.rivm.nl. Accessed 5 Dec 2019.

9. Nic Lochlainn LM, Woudenberg T, van Lier A, Zonnenberg I, Philippi M, de Melker HE, et al. **A novel measles outbreak control strategy in the Netherlands in 2013–2014 using a national electronic immunization register: A study of early MMR uptake and its determinants**. *Vaccine*. 2017, 35:5828–34. doi:10.1016/J.VACCINE.2017.09.018.

10. Ruijs WLML, Hautvast JLLA, Van Der Velden K, De Vos S, Knippenberg H, Hulscher MEEJL. **Religious subgroups influencing vaccination coverage in the Dutch Bible belt: an ecological study**. *BMC Public Health*. 2011, 11:102. doi:10.1186/1471-2458-11-102.

11. Fournet N, Mollema L, Ruijs WL, Harmsen IA, Keck F, Durand JY, et al. **Under-vaccinated groups in Europe and their beliefs, attitudes and reasons for non-vaccination; two systematic reviews**. *BMC Public Health*. 2018, 18:196. doi:10.1186/s12889-018-5103-8.

12. Ruijs WLM, Hautvast JLA, van IJzendoorn G, van Ansem WJC, van der Velden K, Hulscher ME. **How orthodox protestant parents decide on the vaccination of their children: a qualitative study**. *BMC Public Health*. 2012, 12:408. doi:10.1186/1471-2458-12-408.

13. van Lier A, van de Kassteele J, de Hoogh P, Drijfhout I, de Melker H. **Vaccine uptake determinants in The Netherlands**. *Eur J Public Health*. 2014, 24:304–9. doi:10.1093/eurpub/ckt042.

14. Klomp JHE, van Lier A, Ruijs WLM. **Vaccination coverage for measles, mumps and rubella in anthroposophical schools in Gelderland, The Netherlands**. *Eur J Public Health*. 2015, 25:501–5. doi:10.1093/eurpub/cku178.

15. Harmsen IA, Ruiter RAC, Paulussen TGW, Mollema L, Kok G, de Melker HE. **Factors that influence vaccination decision-making by parents who visit an anthroposophical child welfare center: a focus group study.** *Adv Prev Med*. 2012, 2012:175694. doi:10.1155/2012/175694.

16. Bier M, Brak B. **A simple model to quantitatively account for periodic outbreaks of the measles in the Dutch Bible Belt**. *Eur Phys J B*. 2015, 88:107. doi:10.1140/epjb/e2015-50621-9.

17. Klinkenberg D, van Hoek AJ, Veldhuijzen IK, Hahné S, Wallinga J. **Measuring herd protection of unvaccinated children: measles-mumps-rubella vaccination coverage in schools in the Netherlands [Poster]**. In: 7th International Conference on Infectious Disease Dynamics, 3-6 December. 2019.

18. Hagberg AA, Schult DA, Swart PJ. **Exploring network structure, dynamics, and function using NetworkX**. In: 7th Python in Science Conference (SciPy 2008). 2008. p. 11–5. http://conference.scipy.org/proceedings/SciPy2008/paper_2/. Accessed 6 Dec 2019.

19. Diekmann O, Heesterbeek JAP. **Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation**. *Wiley Ser*. 2000, :322. http://eu.wiley.com/WileyCDA/WileyTitle/productCd-0471492418.html.

20. Kucharski AJ, Wenham C, Brownlee P, Racon L, Widmer N, Eames KTD, et al. **Structure and consistency of self-reported social contact networks in British secondary schools**. 2018. doi:10.1371/journal.pone.0200090.

21. Guclu H, Read J, Vukotich CJ, Galloway DD, Gao H, Rainey JJ, et al. **Social Contact Networks and Mixing among Students in K-12 Schools in Pittsburgh, PA**. *PLoS One*. 2016, 11:e0151139.

doi:10.1371/journal.pone.0151139.

22. Grantz, H. K, Cummings DAT, Zimmer SM, Vukotich CJ, Galloway DD, Schweizer M Lou, et al. **Age-specific social mixing of school-aged children in a US setting using proximity detecting sensors and contact surveys**. 2020. doi:10.1101/2020.07.12.20151696.

23. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. **Social contacts and mixing patterns relevant to the spread of infectious diseases.** *PLoS Med*. 2008, 5:e74. doi:10.1371/journal.pmed.0050074.

24. Danon L, Read JM, House TA, Vernon MC, Keeling MJ. **Social encounter networks: characterizing Great Britain.** *Proc Biol Sci*. 2013, 280:20131037. doi:10.1098/rspb.2013.1037.

25. Eames KTD, Tilston NL, Edmunds WJ. **The impact of school holidays on the social mixing patterns of school children**. *Epidemics*. 2011, 3:103–8. doi:10.1016/j.epidem.2011.03.003.

26. le Polain de Waroux O, Saliba V, Cottrell S, Young N, Perry M, Bukasa A, et al. **Summer music and arts festivals as hot spots for measles transmission: experience from England and Wales, June to October 2016**. *Eurosurveillance*. 2016, 21:30390. doi:10.2807/1560-7917.ES.2016.21.44.30390.

9 Discussion

In this thesis I have focused on differences in transmission and vaccination uptake between social groups, and how these may contribute to observed and previously unexplained transmission dynamics in diverse populations. I have approached this problem from a number of angles through a combination of data analysis and simulation studies that use mechanistic models that account for transmission within and between particular social groups within a population.

9.1 Summary of key results

Previously there was broad evidence of differences in infectious disease outcomes between social and ethnic groups[1–6] as well as well-known differences in vaccine uptake between socio-religious groups in multiple settings around the world[7]. There was however little understanding of how the differences in disease outcome relate to transmission heterogeneity in the population, and the key differences between religious groups with low vaccination had not been adequately quantified. Although there had been extensive of work using mathematical models to understand heterogeneity in transmission, only two modelling analyses sought to understand inequalities in infectious disease as a result of transmission dynamics [8, 9]. Both of these demonstrated some potential contribution from geographical distribution by social class, to inequalities in rates of infection between socio-economic groups. Moreover, analysis of inequalities in

infectious disease outcome have not been detailed enough to confirm or rule out the role of transmission and control differences as key drivers.

In chapter 2 I started this investigation by evaluating the relative potential contribution of differences in contact rate, susceptibility and vaccine uptake to inequalities in influenza and rubella infection. To assess how inequalities are impacted by vaccination by comparing rates of infection between groups before and after vaccination. I evaluated overall difference in rate of infection as well as specifically in two risk groups (elderly for influenza and women of childbearing age for rubella). I found that prior to vaccination differences in transmission were likely to cause overall inequalities in influenza but not rubella. However, when considering the particular risk group of women of childbearing age, a social group with reduced transmission (contact rate or susceptibility) would have higher risk compared to a social group with higher transmission. In a scenario with sub-optimal vaccination, inequalities in influenza increase; the social group with higher transmission increased resulting with higher relative risk of infection than pre-vaccination. In contrast, the increased risk of rubella in women of childbearing age in the low transmission group was reduced when vaccination was included. Variance based sensitivity analysis showed that for rubella inequalities were much more sensitive to disparities in vaccine uptake than differences in transmission. However, for influenza inequalities were similarly sensitive to both differences in vaccine uptake and transmission rate. These findings highlight that transmission may contribute meaningfully to disparities in infectious diseases but in particular is likely to impact those observed in infections with lower R_0 such as influenza. Infections with higher R_0 such as rubella are highly sensitive to vaccination uptake and hence transmission rates are unlikely to play a role in inequalities when a vaccine is present.

Focusing on influenza and differences in transmission, in chapter 3 I analysed data from the early phase of the UK outbreak of influenza A H1N1 in 2009, where inequalities had been previously identified[10, 11]. By assessing how the ethnic and socio-economic distribution of infection changed over time in Birmingham and London, I observed that the initiation of sustained transmission in both settings coincided with an increase in relative risk of infection in more deprived areas and among South Asians. In the main, the inequalities observed by socio-economic status were in children (less than 19 years of age). These patterns, along with existing knowledge regarding the relative importance of children in transmission and control of influenza[12] provided motivation to investigate school-based social contact networks informed by national datasets.

Chapter 5 sets out two frameworks for evaluating these contact networks. The first evaluates "opportunity for contact" in schools by estimating the rates at which each ethnic group and deprivation quintile attend the schools with each other ethnic group and deprivation quintile. The second uses data on residence or primary-secondary transition to construct a full network of schools, linked through shared homes. The findings show that, in London schools it is South Asian children who have the highest rates of school attendance within the same institutions, 6.7 times higher than expected by proportional mixing. The number of generations of contacts required to reach equal contact rates is also higher for South Asians than other ethnic groups, reflecting the fact that South Asian children are particularly clustered in the school system. This aligns with the observations in chapter 3 that South Asian children were disproportionately affected during the early phase of the 2009 Influenza A H1N1 outbreak in London and Birmingham. The analysis

also highlighted substantial socio-economic segregation, with the most affluent and most deprived quintiles interacting less than 20% of that expected through proportional mixing.

By simulating outbreaks of influenza on a network of schools, I evaluated the naturally occurring variation in incidence at different stages of an epidemic in chapter 6. Although the analysis showed that the structure predicted only negligible inequality in risk over an entire outbreak for realistic values of R_0, there is evidence that in most cases, the initial phase of an outbreak will not be well mixed in the population regardless of where it begins. The ethnic groups prone to the largest inequalities at the initial stages of an outbreak were South Asian and in particular Bangladeshi (Bangladeshi, Indian and Pakistani). There was also some evidence that outbreaks beginning in areas with large South Asian populations are likely to spread to a higher number of schools more quickly. However, it is not possible to quantify how this reflects on the containability of an outbreak without first quantifying the speed and effectiveness with which public health authorities' might be able to respond to an outbreak in a school.

By constructing a similar network of schools in the Netherlands, in chapter 7, I evaluated potential clustering of children in schools associated with low vaccination uptake: the Dutch Reformed and Anthroposophic schools. My analysis of the network revealed higher homophily and shorter relative network distances than expected in both faith denominations. However, the network proximity of Dutch Reformed schools is markedly more pronounced than Anthroposophic schools, potentially resulting in substantial clustering of unvaccinated children, not just within schools but also between them.

Finally, to evaluate whether the clustering of unvaccinated children observed in chapter 7, was sufficient to explain the observed outbreaks in the MMR vaccine era, I used the same network to simulate outbreaks of measles; detailed in chapter 8. Specifically, I evaluated the role of the clustering within and between schools in the large outbreak of measles in 2013-14. Outbreak simulations on the network were able to describe the scale and geographical distribution cases better than an equivalent spatial approximation of interaction (without explicit information about clustering). Additional analysis revealed that outbreaks initiated in Dutch Reformed schools lead to a higher number of schools and children being infected than those initiated in Anthroposophic schools. This aligns with the findings in chapter 7 of higher connectedness between Dutch Reformed schools and shows that the social structure that results from the school system can account for key aspects of measles epidemiology in the Netherlands.

The program of research has, through a set analyses, provided substantial evidence that social groups within a population could have important implications for the epidemiology of infectious diseases both through clustering of unvaccinated children by faithgroup and through segregation of particular social groups through the school system.

9.2 **Strengths and limitations**

A key strength of this research program is that it offers fresh insights regarding interaction between social structure and heterogeneity in transmission and control of infectious diseases from three distinct angles.

Firstly, by conceptualizing the dynamics of two social groups in a minimal way in chapter 1, I was able to clearly evaluate the relative importance of various heterogeneities to observable differences in risk. This provided a clear distinction in the behaviour of transmission-dependent inequalities between diseases with relatively low R_0 (e.g. influenza), which were sensitive to differences in contact rate, and highly transmissible diseases like measles, which were much more sensitive the variation in vaccine uptake.

Although the magnitude of the inequalities produced are not in themselves interpretable, the division between drivers of inequalities in diseases with low and high R_0 is in itself an important strength. This was useful in defining the analyses later in the thesis: Transmission dynamics of influenza and uptake of MMR.

Secondly, by close analysis of case data in the early phases of the Influenza H1N1 outbreak in 2009 I was able to assess, for the first time, how the observed inequalities developed over time. This revealed patterns consistent with concentrations of transmission within particular social groups – particularly groups of individuals of South Asian ethnicity.

Thirdly, the government schools records offer a novel framework to explicitly model transmission between schools, which are known to be important reservoirs of susceptible hosts. The clarity of this framework allows specific evaluation of the impact of the school network structure on transmission of infectious disease and has proven itself useful in two quite different analyses.

This plurality of approaches amounts to a clear demonstration that social groups should continue to be a key focus of infectious disease research.

Details of the limitations of each analysis are discussed at length in the relevent chapter of this thesis. Here I summarise some limitations to the overall work and challenges which limited research opportunities.

A key limitation of the analysis of the Influenza H1N1 outbreak in the UK, was that data access restrictions and computational limitations meant that using this data to fit a mathematical model was not an option as part of this research program. Were these issues to be resolved, there would be an opportunity to combine this data with the analysis of school networks detailed in chapter 6. In turn this would allow quantification of the contribution of the school network to the observed outbreak dynamics. Likewise, if detailed time history of measles cases were available, a full analysis of the contribution of the school network to the scale of measles outbreaks could be evaluated.

The networks used in the analysis later in the thesis approximate transmission between schools, such that an infected school can infect a susceptible school, assuming the outbreak reaches its estimated final size. This approximation neglects more complex transmission between the schools in two key ways:

Firstly, simulations are ‘generation based’ which means that they do not strictly follow the temporal course of an outbreak but rather show how successive schools might be infected with no information about timescale. This precludes evaluating the modelled cases against epidemiological time series.

Secondly, although in my framework an infected school is constrained to infect an adjacent school and recover in the same step, in reality infection may be transmitted in both directions between two schools once both are infected. This may play a role in persistence and final size. For example, for schools that have a small number of cases but no large outbreak initially, there is a finite probability of re-entry of the infection from a school it infected with the few initial cases. The dynamics of this problem cannot be investigated under the framework as it is and introduce heterogeneity in persistence of an outbreak across the network.

In order to include the potential for successive generations of schools to be infected at the same time and affect each other's outbreaks, the model would need to increase significantly in complexity and computational demand. Under this analysis, where assessments are broadly performed either on the properties of individual schools or on an entire outbreak, temporal information is not required. This extension could be made to the modelling framework at a later date if required for future research pursuits.

Another important limitation of the network model framework is that it only considers heterogeneity in transmission introduced through the school system. There may be other factors that affect these communities differently, such as attendance at religious gatherings, summer camps and general behavioural differences. If appropriate data were found to support additional assumptions these factors could be incorporated in an extended model, although complexity would likely increase as school units may not represent appropriate groupings to reflect this additional structure. This would have to be

addressed by either increasing computational resources or reducing the scale of the setting analysed.

A very important limitation of the network models is also the simplification of within-school transmission. There have been a number of excellent studies in schools to measure contact between pupils, which reveal strong clustering within age groups[13–15]. This was not included in the framework in the interest of parsimony. Future analysis could address this limitation by including grade structure within schools, however this would be challenging for large scale networks and would require data on contacts between schools by age group.

Finally, data for explicit household links between schools were used in the model of the Netherlands but these were not available for London. Such data does exist for the UK, but due to arduous application procedures, access was unrealistic within the timeframe of the research program. Instead, I made estimates from available data on transfer of students from primary to secondary schools in London. If the data on the explicit links between schools were available for the whole of the UK, analysis of Birmingham and London would have been preferable to London alone to match the analysis of influenza case data from 2009.

9.3 Contributions of this research relative to previous knowledge

This program of research provides several meaningful steps forwards from previous knowledge.

Firstly, previously there was a general understanding that due to herd effects populations with higher transmission would have reduced protection from vaccination[16]. Chapter 1 extended this by evaluating the impact that might be expected on existing inequalities in disease due to suboptimal vaccination. Moreover, identifying that vaccine uptake dominates inequalities in highly transmissible infections, whereas transmission differences are more likely to play a part is infections with lower R_0 is an important finding that helps to understand where observed inequalities in vaccine-controlled infections is likely to originate.

Secondly, inequalities in influenza incidence in the UK had only been observed over large geographical regions and for the early phase of the 2009 outbreak[10, 11]. By closer analysis of this data I have provided important insight into how those inequalities developed over time. This supports understanding as to whether the inequalities are differences in disease or something else unconnected to transmission, such as reporting or severity. Importantly, the observation that inequality in disease may be present early in an outbreak but with no particular social group at structurally higher risk of these inequalities highlights the importance of evaluating inequalities even when majority groups are disproportionately affected as well as when minority groups are, rather than selectively reporting higher incidence in particular groups of interest.

Thirdly, the work offers important developments in understanding measles dynamics in the Netherlands. I was able to show explicitly the relative proximity of Dutch Reformed schools. This leads to clustering of unvaccinated children across the Netherlands, which provides a basis for the large outbreaks of measles [17]. The framework can now be used to provide a better indication of risk at the early phases of an outbreak.

Finally, developing a modelling framework that effectively captures interaction between social groups in school-aged children is possibly the most important contribution of this work. Having a framework which is able to capture these dynamics in a parsimonious way is an important step towards understanding disease transmission dynamics between social groups in a population. Previous models[8, 9] have relied on detailed synthetic populations which, whilst able to capture multiple nuances in population structure, can provide challenges in interpretation. Furthermore, without specific parameterisation from similar data, they may not capture specific properties of the school network offered by the approach taken here.

9.4 Implications and future research opportunities

This work has some important implications for the understanding of the potential impact of social structure on inequalities in risk and control of infectious disease. There are three particularly clear points:

Firstly, the results shown in Chapter 2 have important implications regarding how observed inequalities in childhood infections should be interpreted. Concretely, inequalities in diseases with a widely distributed vaccine program in place, are likely to be a result of variation in uptake as opposed to variation in transmission. However, this may not be the case for inequalities during outbreaks of influenza. Moreover, in a scenario where there are inequalities in transmission of influenza, suboptimal vaccine coverage may accentuate these disparities if not properly accounted for.

Secondly, Inequalities in influenza outbreaks are complex, single estimates of disparities in risk over a fixed time may be misleading. Importantly the dynamics of infection on the school network revealed that observations of inequalities were highly likely early in an outbreak. This coincides with the period when inequalities can be measured most precisely, as highly detailed data is often only available when case numbers are low. Although outbreaks which originate in South Asian communities showed more severe inequalities than among other ethnicities, and therefore may be more readily observed. Simulations showed that on average the network preferred higher risk in more affluent white children at early stages. This may point towards bias in reporting inequalities in marginalised groups over majority groups, this could explain the persistence of reports of higher risk within "ethnic minorities" in multiple countries[18–26] with no clear mechanism to link them, whilst there is also evidence that this effect is not systematic at all[27].

Thirdly, my finding that the extent of measles outbreaks in the Netherlands can be explained by school and household transmission helps to elucidate a previously poorly understood phenomenon where a population with high vaccination uptake continues to sustain large outbreaks. Furthermore, the framework provides a means to evaluate the potential for an outbreak to reach large numbers of unvaccinated children. For example the difference between outbreaks originating in Anthroposophic and Dutch Reformed schools. Although the revelations do not provide a clear means to control these outbreaks, the support in understanding risk could aid proportionate and targeted response.

Although this framework could, in principal, be applied to other countries, care should be taken when generalising the particular findings of this work. The results suggest that

concentration of unvaccinated children within particular schools can result in large clusters of unvaccinated children across a country. However, this probably needs to be replicated at multiple levels of education (e.g. primary and secondary schools) for this phenomenon to have full effect. Moreover, it is not uncommon for children in the Netherlands to travel between cities for secondary school, this is possibly due to an extraordinarily high population density and may not be replicated in other countries with sparser populations.

The framework developed has the potential to provide a basis for analysis beyond the focuses of this thesis. For example, the network provides a natural tool for assessing the impact of school closures as an intervention to control outbreaks, the impact of making immunization mandatory within particular schools and assessment of spatial risk when outbreaks are detected in certain institutions. This work could easily be extended to multiple pathogens in multiple locations.

An important next step is to validate the assumption in the model that transmission through households represents a good indicator of overall risk of transmission between groups. One approach may be to sequence influenza samples from infected school children and evaluate the relationship between network distance and phylogeny i.e. are closer strains more proximal on the network.

This work was focused on the role of the network in particular epidemiological phenomena that can be observed. There is however, opportunity to explore more general properties of the network, which may provide information about transmission dynamics of childhood infections which prove useful for control (e.g. the relative infectiousness of

primary and secondary schools). If data were to be made available for the entire country, there would be an opportunity to evaluate the differences in network structure between settings. This may be useful in understanding the potential for an outbreak to spread between towns and cities.

Finally, a clear opportunity for future research is evaluation of the difference in expected transmission dynamics and resultant epidemiology of particular pathogens between countries as a result of their different school systems. The United States for example have a much more regulated state school system than the UK or the Netherlands, as children are required to attend schools within their particular district[28]. This may to lead to much different dynamics from those seen in the UK and Netherlands, where families are offered more choice when selecting a secondary school. These fundamental differences in structure may also impact the effectiveness of school-based interventions, such as school closures or school-based vaccination programs.

Social structure as a result of preferential contact within particular social groups, has demonstrable impact on the dynamics and control of infectious diseases. The effects of these heterogeneities are complicated and differ between settings and pathogens. The work in this thesis provides insight regarding these dynamics and provides a new framework for evaluating them within school-age children, a key population in the epidemiology of many acute infections. My hope is that this work can be developed and extended to provide further insights that can in turn support public health policy in the future and ultimately improve the control of infectious disease both in effectiveness and equity.

9.5 References

1. Semenza JC, Suk JE, Tsolova S. **Social determinants of infectious diseases: a public health priority.** *Euro Surveill*. 2010, 15:2–4. doi:10.2807/ese.15.27.19608-en.

2. Semenza JC, Giesecke J. **Intervening to Reduce Inequalities in Infections in Europe**. *Am J Public Health*. 2008, 98:787–92. doi:10.2105/AJPH.2007.120329.

3. Semenza JC. **Strategies to intervene on social determinants of infectious diseases**. *Eurosurveillance*. 2010, 15:32–9. doi:10.2807/ese.15.27.19611-en.

4. Hawker J, Olowokure B, Sufi F, Weinberg J, Gill N, Wilson RC. **Social deprivation and hospital admission for respiratory infection:**. *Respir Med*. 2003, 97:1219–24. doi:10.1016/S0954-6111(03)00252-X.

5. Myles PR, McKeever TM, Pogson Z, Smith CJP, Hubbard RB. **The incidence of pneumonia using data from a computerized general practice database.** *Epidemiol Infect*. 2009, 137:709–16. doi:10.1017/S0950268808001428.

6. Pockett RD, Adlard N, Carroll S, Rajoriya F. **Paediatric hospital admissions for rotavirus gastroenteritis and infectious gastroenteritis of all causes in England: an analysis of correlation with deprivation.** *Curr Med Res Opin*. 2011, 27:777–84. doi:10.1185/03007995.2011.555757.

7. Fournet N, Mollema L, Ruijs WL, Harmsen IA, Keck F, Durand JY, et al. **Under-vaccinated groups in Europe and their beliefs, attitudes and reasons for non-vaccination; two systematic reviews**. *BMC Public Health*. 2018, 18:196. doi:10.1186/s12889-018-5103-8.

8. Hyder A, Leung B. **Social deprivation and burden of influenza: Testing hypotheses and gaining insights from a simulation model for the spread of influenza**. *Epidemics*. 2015, 11:71–9. doi:10.1016/j.epidem.2015.03.004.

9. Kumar S, Piper K, Galloway DD, Hadler JL, Grefenstette JJ. **Is population structure sufficient to generate area-level inequalities in influenza rates? An examination using agent-based models**. *BMC Public Health*. 2015, 15:947. doi:10.1186/s12889-015-2284-2.

10. Balasegaram S, Ogilvie F, Glasswell A, Anderson C, Cleary V, Turbitt D, et al. **Patterns of early transmission of pandemic influenza in London - link with deprivation.** *Influenza Other Respi Viruses*. 2012, 6:e35-41. doi:10.1111/j.1750-2659.2011.00327.x.

11. Inglis NJ, Bagnall H, Janmohamed K, Suleman S, Awofisayo A, De Souza V, et al. **Measuring the effect of influenza A(H1N1)pdm09: the epidemiological experience in the West Midlands, England during the "containment" phase.** *Epidemiol Infect*. 2014, 142:428–37. doi:10.1017/S0950268813001234.

12. Baguelin M, Flasche S, Camacho A, Demiris N, Miller E, Edmunds WJ. **Assessing Optimal Target**

Populations for Influenza Vaccination Programmes: An Evidence Synthesis and Modelling Study. *PLoS Med*. 2013, 10:e1001527. doi:10.1371/journal.pmed.1001527.

13. Kucharski AJ, Wenham C, Brownlee P, Racon L, Widmer N, Eames KTD, et al. **Structure and consistency of self-reported social contact networks in British secondary schools**. *PLoS One*. 2018, 13:e0200090. doi:10.1371/journal.pone.0200090.

14. Grantz, H. K, Cummings DAT, Zimmer SM, Vukotich CJ, Galloway DD, Schweizer M Lou, et al. **Age-specific social mixing of school-aged children in a US setting using proximity detecting sensors and contact surveys**. 2020. doi:10.1101/2020.07.12.20151696.

15. Guclu H, Read J, Vukotich CJ, Galloway DD, Gao H, Rainey JJ, et al. **Social Contact Networks and Mixing among Students in K-12 Schools in Pittsburgh, PA**. *PLoS One*. 2016, 11:e0151139. doi:10.1371/journal.pone.0151139.

16. Fine P, Eames K, Heymann DL. **"Herd Immunity": A Rough Guide**. *Clin Infect Dis*. 2011, 52:911–6. doi:10.1093/cid/cir007.

17. Woudenberg T, van Binnendijk RS, Sanders EAM, Wallinga J, de Melker HE, Ruijs WLM, et al. **Large measles epidemic in the Netherlands, May 2013 to March 2014: changing epidemiology.** *Euro Surveill*. 2017, 22. doi:10.2807/1560-7917.ES.2017.22.3.30443.

18. Charland KM, Brownstein JS, Verma A, Brien S, Buckeridge DL. **Socio-Economic Disparities in the Burden of Seasonal Influenza: The Effect of Social and Material Deprivation on Rates of Influenza Infection**. *PLoS One*. 2011, 6:e17207. doi:10.1371/journal.pone.0017207.

19. Dee DL, Bensyl DM, Gindler J, Truman BI, Allen BG, D'Mello T, et al. **Racial and Ethnic Disparities in Hospitalizations and Deaths Associated with 2009 Pandemic Influenza A (H1N1) Virus Infections in the United States**. *Ann Epidemiol*. 2011, 21:623–30. doi:10.1016/j.annepidem.2011.03.002.

20. Quinn SC, Kumar S, Freimuth VS, Musa D, Casteneda-Angarita N, Kidwell K. **Racial Disparities in Exposure, Susceptibility, and Access to Health Care in the US H1N1 Influenza Pandemic**. *Am J Public Health*. 2011, 101:285–93. doi:10.2105/AJPH.2009.188029.

21. Levy NS, Nguyen TQ, Westheimer E, Layton M. **Disparities in the Severity of Influenza Illness**. *J Public Heal Manag Pract*. 2013, 19:16–24. doi:10.1097/PHH.0b013e31824155a2.

22. Navaranjan D, Rosella LC, Kwong JC, Campitelli M, Crowcroft N. **Ethnic disparities in acquiring 2009 pandemic H1N1 influenza: a case–control study**. *BMC Public Health*. 2014, 14:214. doi:10.1186/1471-2458-14-214.

23. Placzek H, Madoff L. **Effect of Race/Ethnicity and Socioeconomic Status on Pandemic H1N1-Related Outcomes in Massachusetts**. *Am J Public Health*. 2014, 104:e31–8. doi:10.2105/AJPH.2013.301626.

24. Mayoral JM, Alonso J, Garín O, Herrador Z, Astray J, Baricot M, et al. **Social factors related to the clinical severity of influenza cases in Spain during the A (H1N1) 2009 virus pandemic**. *BMC Public Health*. 2013, 13:118. doi:10.1186/1471-2458-13-118.

25. Wilson N, Barnard LT, Summers JA, Shanks GD, Baker MG. **Differential Mortality Rates by Ethnicity in 3 Influenza Pandemics Over a Century, New Zealand**. *Emerg Infect Dis*. 2012, 18:71–7. doi:10.3201/eid1801.110035.

26. Yousey-Hindes KM, Hadler JL. **Neighborhood Socioeconomic Status and Influenza**

Hospitalizations Among Children: New Haven County, Connecticut, 2003–2010. *Am J Public Health*. 2011, 101:1785–9. doi:10.2105/AJPH.2011.300224.

27. Tricco AC, Lillie E, Soobiah C, Perrier L, Straus SE. **Impact of H1N1 on socially disadvantaged populations: Summary of a systematic review**. *Influenza Other Respi Viruses*. 2013, 7 SUPPL.2:54–8. doi:10.1111/irv.12082.

28. Snyder TD, de Brey C, Dillow S. **Digest of Education Statistics, 2017**.

Appendix A. Suplementary material for Analysis A

Additional file 1: Quantifying the impact of social groups and vaccination on inequalities in infectious diseases using a mathematical model

Contents:

Mathematical model

Model Structure

We developed a Susceptible Exposed Infected Recovered (SEIR) model with two social groups and 15 age groups. The full system of ordinary differential equations is expressed as:

$$D\left(C_{i,G}\right)$$

$$= \begin{cases} (1-\rho)\mu - C_{i,G}a_i & ; \quad i = 1, C = S \\ \rho\mu - C_{i,G}a_i & ; \quad i = 1, C = V \\ -C_{i,G}a_i & ; \quad i = 1, C \neq S, V \\ C_{i-1,G}a_{i-1} - C_{i,G}a_i & ; \quad 1 < i < n \\ C_{i-1,G}a_{i-1} - \dfrac{C_{j,G}\mu}{p_{i,G}} & ; \quad i = n \end{cases} \quad ; \quad C_{i,G} \in \left\{S_{i,G}, E_{i,G}, I_{i,G}, R_{i,G}, V_{i.G}\right\}$$

$$p_{i,G} = S_{i,G} + E_{i,G} + I_{i,G} + R_{i,G} + V_{i.G}$$

$$\lambda_{i,H} = \sum_{j=1}^{n_{age}} \beta_{ij}\left(I_{j,A} + \xi I_{j,B}\right)$$

$$\lambda_{i,L} = \sum_{j=1}^{n_{age}} \eta\beta_{ij}\left(\xi I_{j,A} + \chi I_{j,B}\right)$$

$$\frac{dS_{i,G}}{dt} = -\sum_{j=1}^{n} \lambda_{i,G}\, S_{i,G} - \mu S_{i,G} + D\left(S_{i,G}\right)$$

$$\frac{dE_{i,G}}{dt} = \sum_{j=1}^{n} \lambda_{i,G}\, S_{i,G} - \sigma E_{i,G} + D\left(E_{i,G}\right)$$

$$\frac{dI_{i,G}}{dt} = \sigma I_{i,G} - rI_{i,G} + D\left(I_{i,G}\right)$$

$$\frac{dR_{i,G}}{dt} = \kappa R_{i,G} + D\left(R_{i,G}\right)$$

$$\frac{dV_{i,G}}{dt} = D\left(V_{i,G}\right),$$

where $S_{i,G}$, $E_{i,G}$, $I_{i,G}$, $R_{i,G}$, $V_{i.G}$ are the proportion of the population in age group i and in social group G that are susceptible, exposed (infected but not infectious), infectious, recovered and vaccinated respectively. We define the parameters:

μ is birth and death rate

ρ is the proportion of the population vaccinated

β_{ij} is the rate of transmission from age group j to age group i

η is the relative susceptibility of group L

χ is the relative contact rate of group L

ξ is the rate of contact between social groups relative to within group H

$p_{i,G}$ is the proportion of the population in age group i and social group G \in $\{H, L\}$

σ is the rate at which individuals become infectious after being infected

κ is the rate at which individuals recover from infection (cease to be infectious)

a_i is the rate at which the population moves from age group i to age group i+1

Calculating the force of infection, λ

To ensure that all parameterisations of differences in contact, susceptibility and social integration result comparable epidemiology, we kept the basic reproduction number (R_0) constant by scaling the next generation matrix **R** linearly such that its largest eigenvalue was equal to the correct value of R_0.

Each element of the next generation matrix, R_{ab}, gives the expected number of cases in age and social group a resulting in transmission from a single case in age and social group b in an otherwise totally susceptible population. The force

of infection vector can be written in terms of the next generation matrix R, which is a function of the matrix of transmission parameters, **B** and the infectious period γ.

$$\vec{\lambda} = \mathbf{B}\vec{\mathrm{I}} = \frac{1}{\gamma}\mathbf{R}\vec{\mathrm{I}}$$

Neglecting age groups initially, we rewrite the force of infection as a function of the transmission rate within age group H, β and the social interaction between the two social groups, **X**:

$$\vec{\lambda} = \beta\mathbf{X}\vec{\mathrm{I}},$$

where,

$$\begin{pmatrix}\lambda_{\mathrm{H}}\\ \lambda_{\mathrm{L}}\end{pmatrix} = \beta\begin{pmatrix}1 & \xi\\ \eta\xi & \eta\chi\end{pmatrix}\begin{pmatrix}\mathrm{I}_{\mathrm{H}}\\ \mathrm{I}_{\mathrm{L}}\end{pmatrix}$$

We introduced age groups by first defining **P** as a normalised age dependent social contact matrix (such that the elements sum to unity) where each element, P_{ij}, is the rate of contact between age group *j* and age group *i* per individual in group *i.* We construct a matrix, $\mathbf{P_{total}}$, to account for the transmission between age groups and between two social groups:

$$\mathbf{P_{total}} = \begin{pmatrix}\mathbf{P} & \mathbf{P}\\ \mathbf{P} & \mathbf{P}\end{pmatrix}$$

.

We take the Hadamard (element-by-element) product of **X** and $\mathbf{P_{total}}$ to give a normalised age and social group dependent next generation matrix, $\mathbf{R}_{norm}$.

$$\mathbf{R}_{norm} = \mathrm{X} \circ \mathbf{P_{total}} = \begin{pmatrix}1 & \xi\\ \eta\xi & \eta\chi\end{pmatrix} \circ \begin{pmatrix}\mathbf{P} & \mathbf{P}\\ \mathbf{P} & \mathbf{P}\end{pmatrix} = \begin{pmatrix}\mathbf{P} & \xi\mathbf{P}\\ \eta\xi\mathbf{P} & \eta\chi\mathbf{P}\end{pmatrix}$$

The next generation matrix is proportional to this normalised matrix,

$$\mathbf{R} = r\mathbf{R}_{norm}$$

We select a value of *r* to give a fixed R_0, defined as the spectral radius of $\mathbf{R}$.

An alternative approach is to vary the ratio of total contact rate in group L and group H by adapting the parameterisation to the following form.

$$\mathbf{R}_{norm} = \mathrm{X} \circ \mathbf{P_{total}} = \begin{pmatrix}1 & \xi\\ \eta\xi & \eta\chi'\end{pmatrix} \circ \begin{pmatrix}\mathbf{P} & \mathbf{P}\\ \mathbf{P} & \mathbf{P}\end{pmatrix} = \begin{pmatrix}\mathbf{P} & \xi\mathbf{P}\\ \eta\xi\mathbf{P} & \eta\chi'\mathbf{P}\end{pmatrix}$$

Where, $\chi' = \chi(\chi - 1)\xi$. The two approaches were found to have consistent results (results not shown).

Parameterisation of the model

Integration of Social-groups

To estimate the proportion of contacts occurring between social groups, we use data collected as part of the Great Britain arm of the POLYMOD survey[1]. Participants were asked to complete a diary of social contacts made over a 24-hour period. As part of the survey, participants recorded the location of each contact event (home, work, school, leisure, transport or other place) and the duration of each contact. For each participant we assume that contact events that occurred within their home were with a member of their own social group. The mean number and time spent with household contacts account for 43% of contact events and for 47% of the total duration of contact the participants report, respectively. In addition, we assume 70–90% of contact that occurred outside the participant's home was also with members of the participant's own social group. The non-household contacts in the same social group then accounted for a mean of 44–56% of contact events and 37–48% of total duration of contact. We assume all remaining 10– 30% of non-household contacts belong to the other social group. To calculate the value of the integration parameter, ξ, we use the ratio between the contact with other social groups and with the participant's social group for each participant in the data set. As the mean ratio was between 0.06 and 0.22 for number of contacts and 0.06 and 0.20 for total duration of contacts (Figure 2), we set ξ between 0.05 and 0.25.

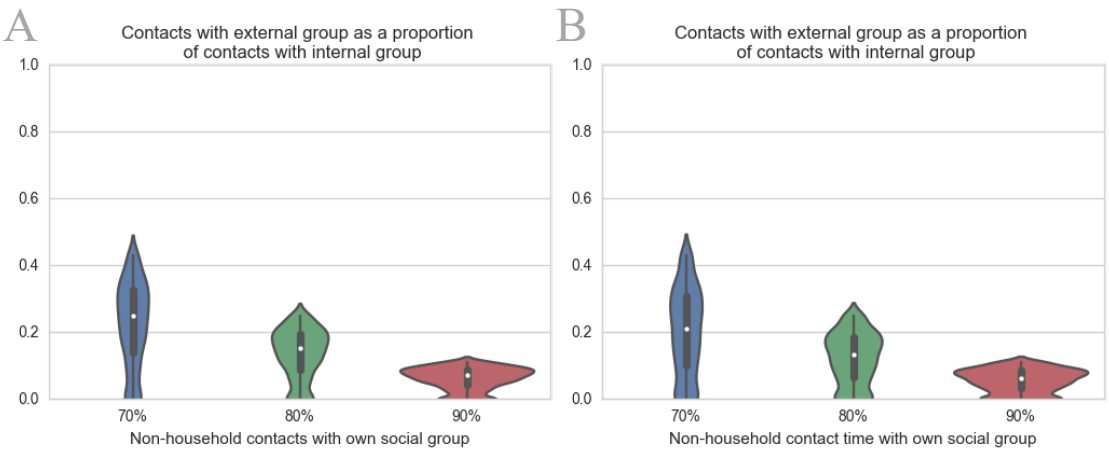


Figure S1 Distribution of the ratio of A) number and B) total duration of contacts within the participants social group and outside of the participants social group (using the GB arm of the POLYMOD contact survey data).

Difference in rate of contact between groups

We also estimated an appropriate range for the difference in rate of contact between groups by also analysing data collected as part of the POLYMOD survey. Firstly, we calculated the number of contacts from the same social group, under the same definition as for integration of social groups.

As part of the POLYMOD survey household size of the participants was recorded. In this data and other contact surveys, household size has been shown to be a good predictor of contact rate, with higher rates in members of larger households[1–7]. In addition, household size distribution can vary significantly between social and ethnic groups. In order to choose an appropriate range for the difference in rate of contact between groups, we created 5 subsets of the sample population by sorting by Household size. We first stratified the population into 16 5-year age groups. To ensure the age distribution of each subset remained the same, subsequently we stratified each age group into 5 quintiles based on household size. We then assemble five final quintiles that comprise similar age distributions (Figure 2) but differing household size distributions (Figure 3). Consequently, the distribution of number of contacts and total duration of contact is also different for each of the quintiles (Figure 4 and Figure 5).

We calculate the ratio of mean number of contacts for each pair of quintiles to give a range of values for the relative rates of contact between two social groups (Figure 7). We use this range as our relative difference in within-group contact parameter (χ = 0.65 – 0.95).

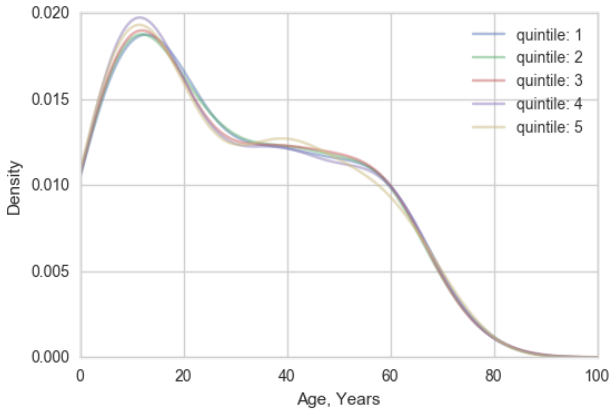


Figure S2 The age distribution of each household size quintile

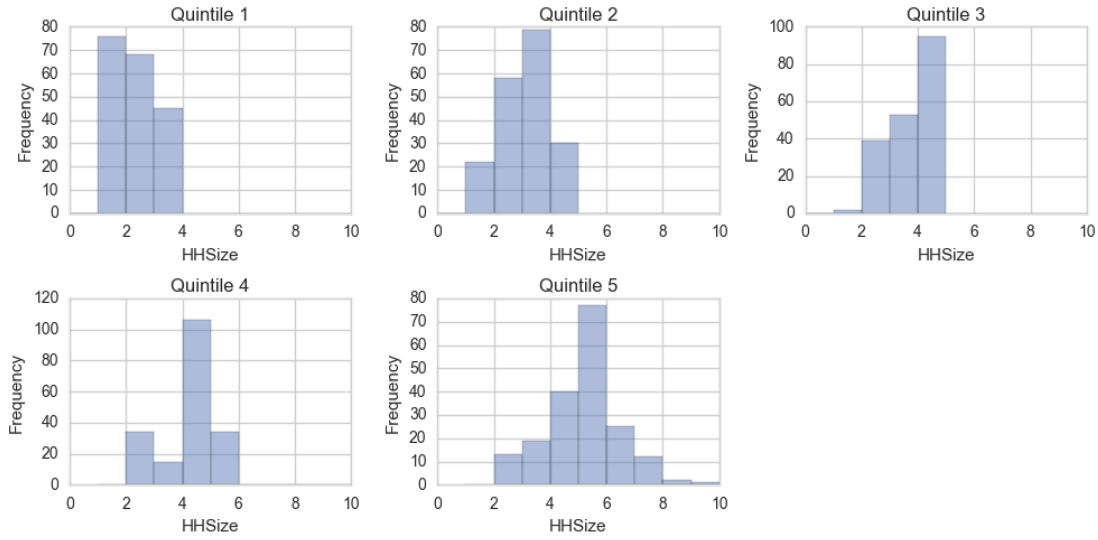


Figure S3 Household size distributions of quintiles across all ages

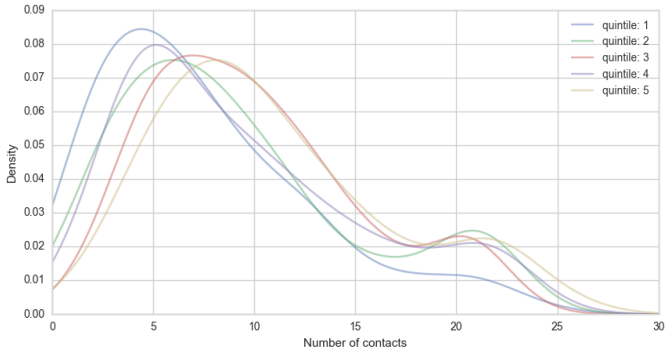


Figure S4 Kernel density estimate showing the distribution of number of contacts in each quintile

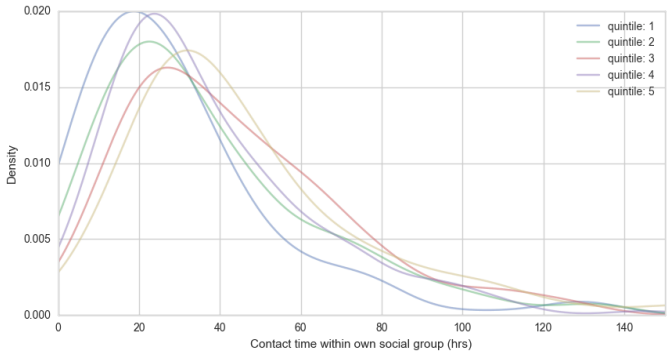


Figure S5 Kernel density estimate showing the distribution of total duration of contacts in each quintile

A B

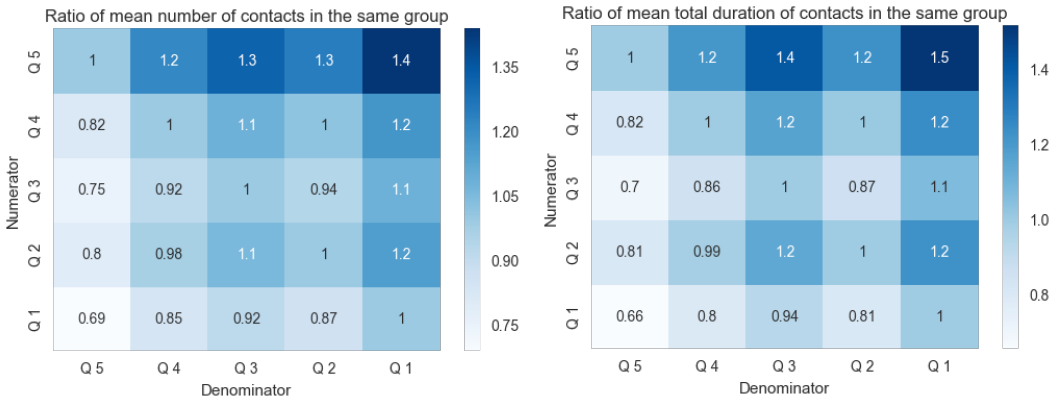


Figure S6 Colour maps showing the ratio of A) number of contacts and B) total duration of contacts between each of the quintiles of POLYMOD participants stratified by household size. The range of the ratios was subsequently used to inform the contact rate in group L relative to group H.

Epidemiological results over the full parameter ranges

To measure the relative risk of infection resulting from a range of differences in contact rate, susceptibility and integration of social groups, we model an outbreak of influenza and endemic rubella in two social groups. We alter relative contact rates within social groups by varying a parameter, χ, relative susceptibility in the social groups by varying a second parameter, η, and the level of integration of the two social groups with each other by varying another parameter, ξ. Figures in the main text only contain results for $\xi = 0.15$. This section presents figures for results with different values of ξ.

Figure 7 to Figure 12 show the cumulative incidence of influenza over an outbreak and cumulative risk of infection for rubella over the first 75 years of life. Each shows the result with no difference between the social groups ($\chi, \eta = 1$) and the range of results in each social group over the full range of either relative contact rate, $\chi = 0.65 - 0.95$ (Figure 7 – Figure 9), or relative susceptibility, $\eta = 0.65 - 0.95$ (Figure 10 – Figure 12), with for a fixed value of $\xi \in \{0.05,\ 0.15,\ 0.25\}$.

Adding differences in contact rate and susceptibility between the two social groups changes the epidemiology and the risk of infection: specifically the risk of infection with influenza increases in group H and reduces in group L. When we increase the integration between the social groups (ξ), there is a reduction in the change in risk relative to when the subgroups are identical. When vaccination is introduced into the model at 80% of the critical vaccination threshold, The epidemiology of both infections changed markedly with overall reduction of risk of infection in both groups. However, indirect protection is greater in the social group L; the with lower rates of transmission (contact rate or susceptibility). This difference in a greater reduction in disease risk than in group H. The consequence is an increase

in the relative risk of infection in influenza in group H and a greater risk of infection from rubella in group H than group L (Figure 13).

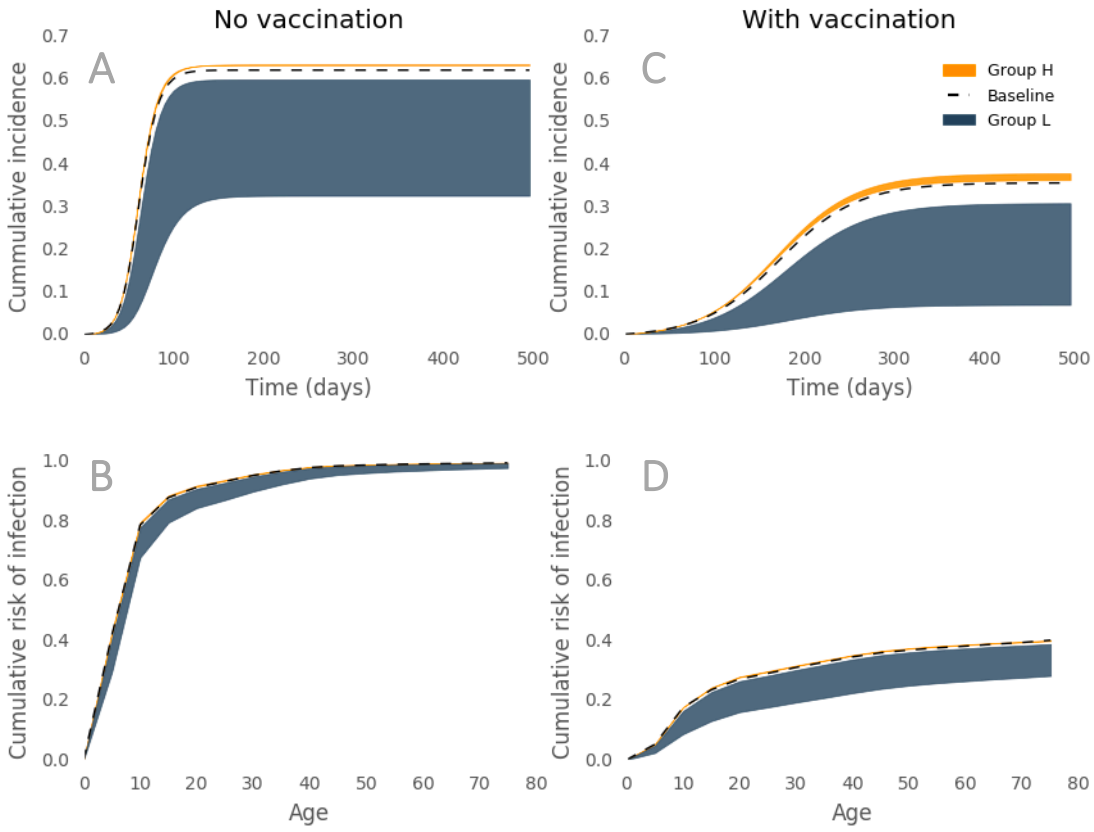


Figure S7 The change in infection over time or by age predicted by the model for seasonal influenza and rubella. With no differences between two population groups under baseline parameters shown in black dashed line and with differences in susceptibility and contact rate for group H in the orange region and group L in the navy region. The results are based on a ratio of contact rate between groups ($\chi = 0.65 – 0.95$) and low integration ($\xi = 0.05$) . A) the cumulative incidence of influenza over a single outbreak with no vaccination, B) shows the proportion of population infected with Rubella by age at endemic equilibrium with no vaccination, C) the cumulative incidence of influenza in remaining unvaccinated individuals with 37% vaccine uptake (80% of the critical vaccination threshold) and D) the proportion of remaining unvaccinated population infected with Rubella by age with 67% vaccine uptake (80% of the critical vaccination threshold).

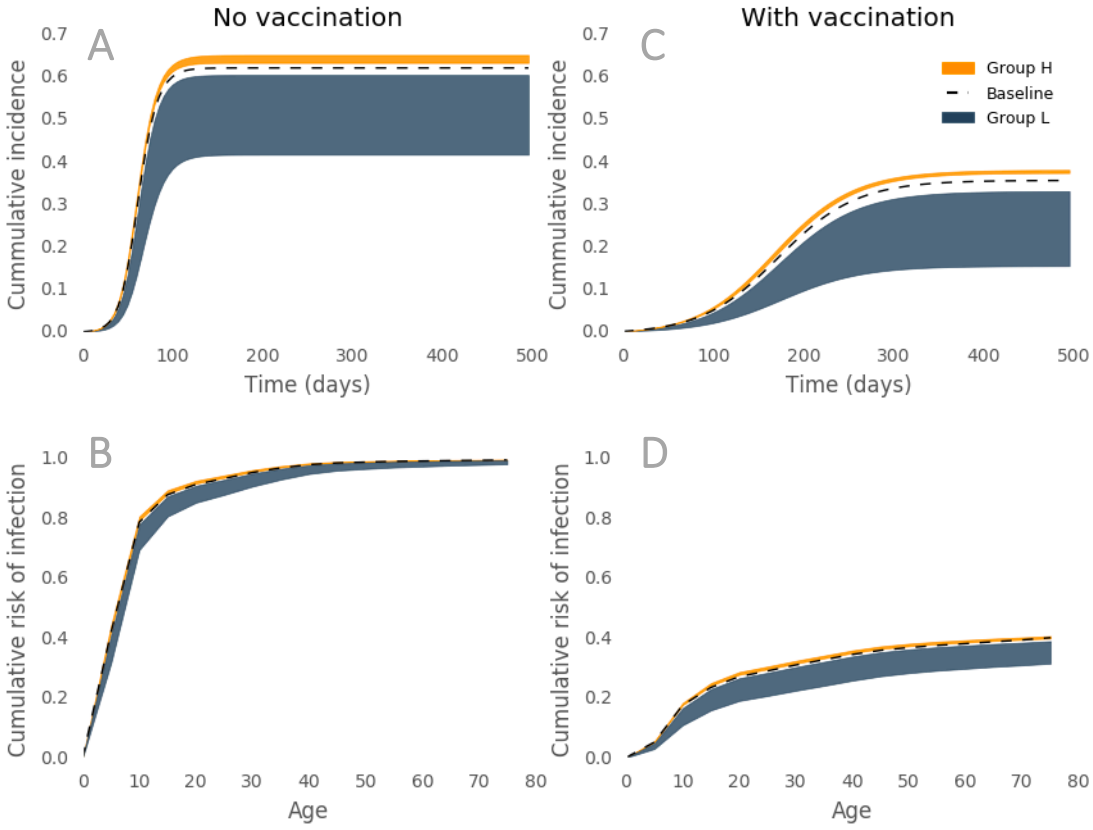


Figure S8 As 7 but with intermediate integration ($\xi = 0.15$)

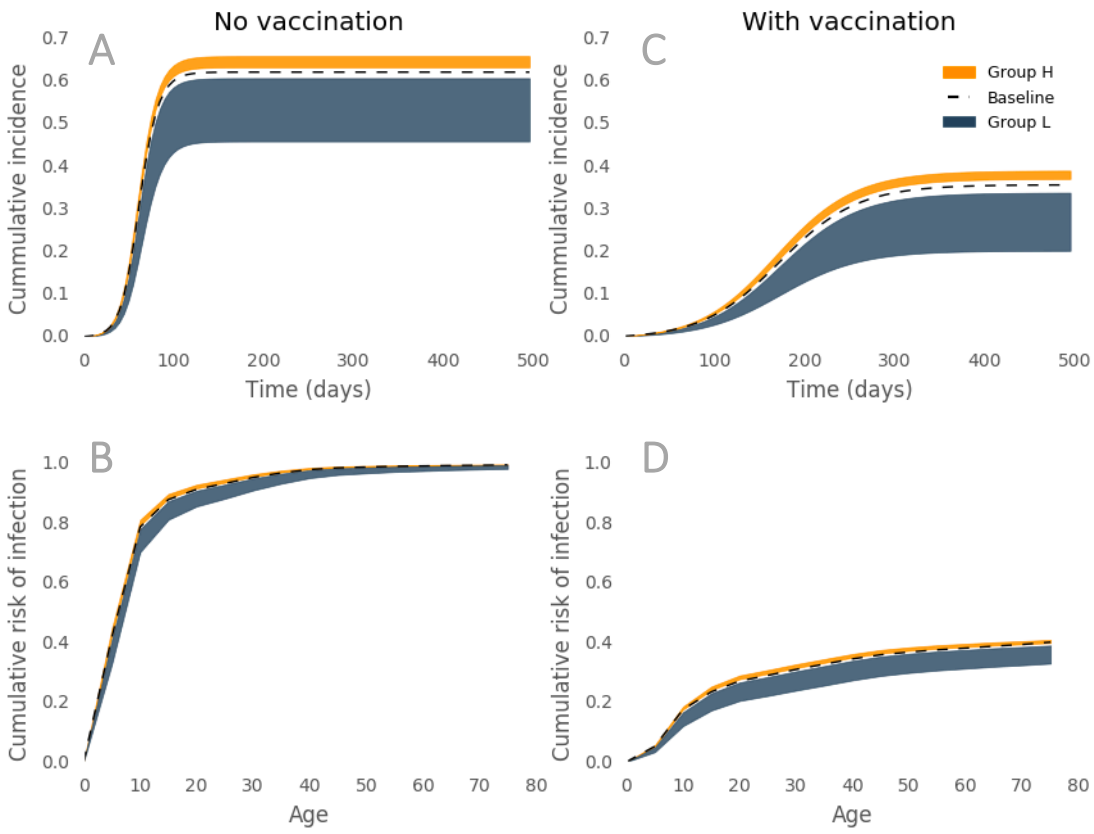


Figure S9 As 7 but with high integration ($\xi = 0.25$)

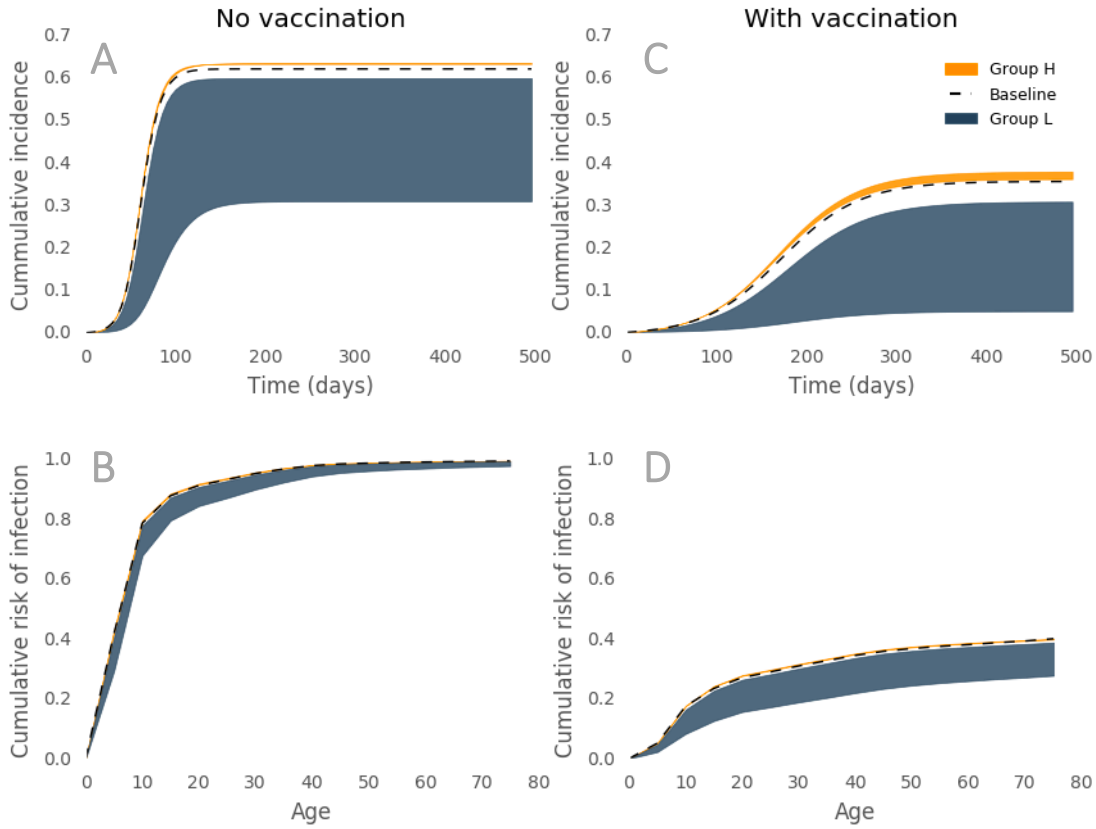


Figure S10 The change in infection over time or by age predicted by the model for seasonal influenza and rubella. With no differences between two population groups under baseline parameters shown in black dashed line and with differences in susceptibility and contact rate for group H in the orange region and group L in the navy region. The results are based on a ratio of susceptibility between groups ($\eta = 0.65 – 0.95$) and low integration ($\xi = 0.05$). A) the cumulative incidence of influenza over a single outbreak with no vaccination, B) the proportion of population infected with Rubella by age at endemic equilibrium with no vaccination, C) the cumulative incidence of influenza in remaining unvaccinated individuals with 37% vaccine uptake (80% of the critical vaccination threshold) and D) the proportion of remaining unvaccinated population infected with Rubella by age with 67% vaccine uptake (80% of the critical vaccination threshold).

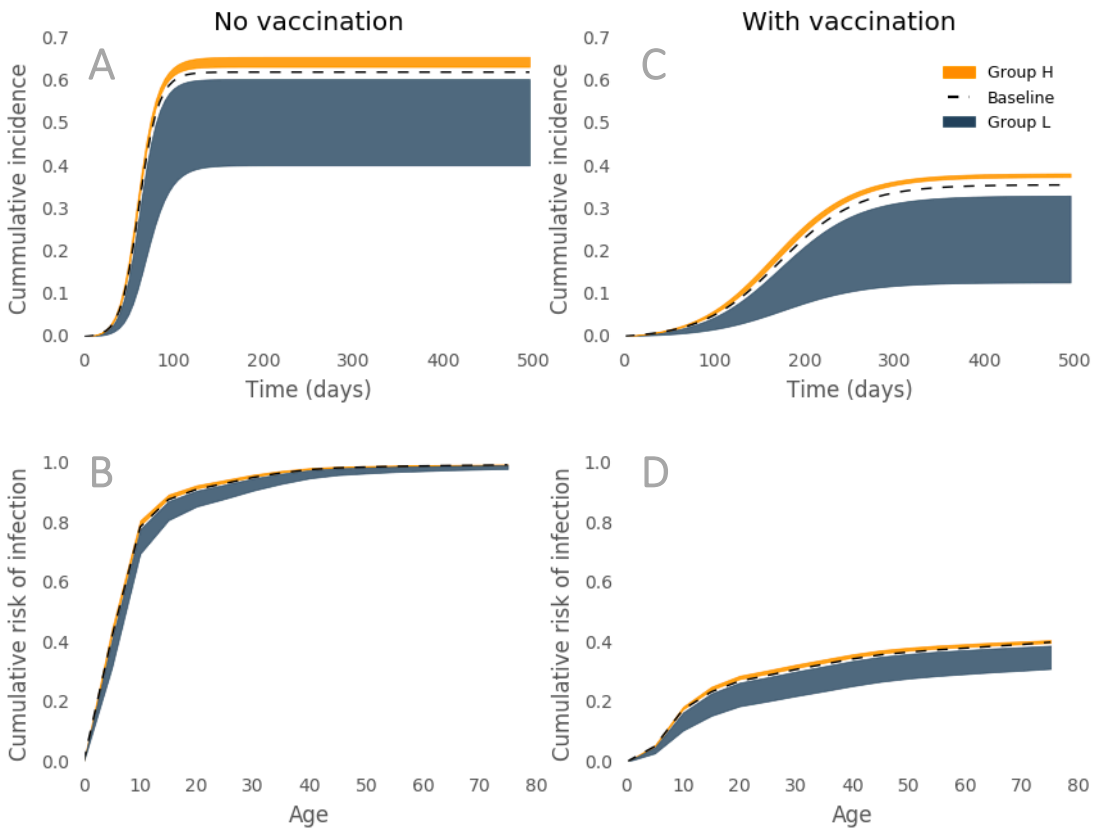


Figure S11 As 10 but with intermediate integration ($\xi = 0.15$)

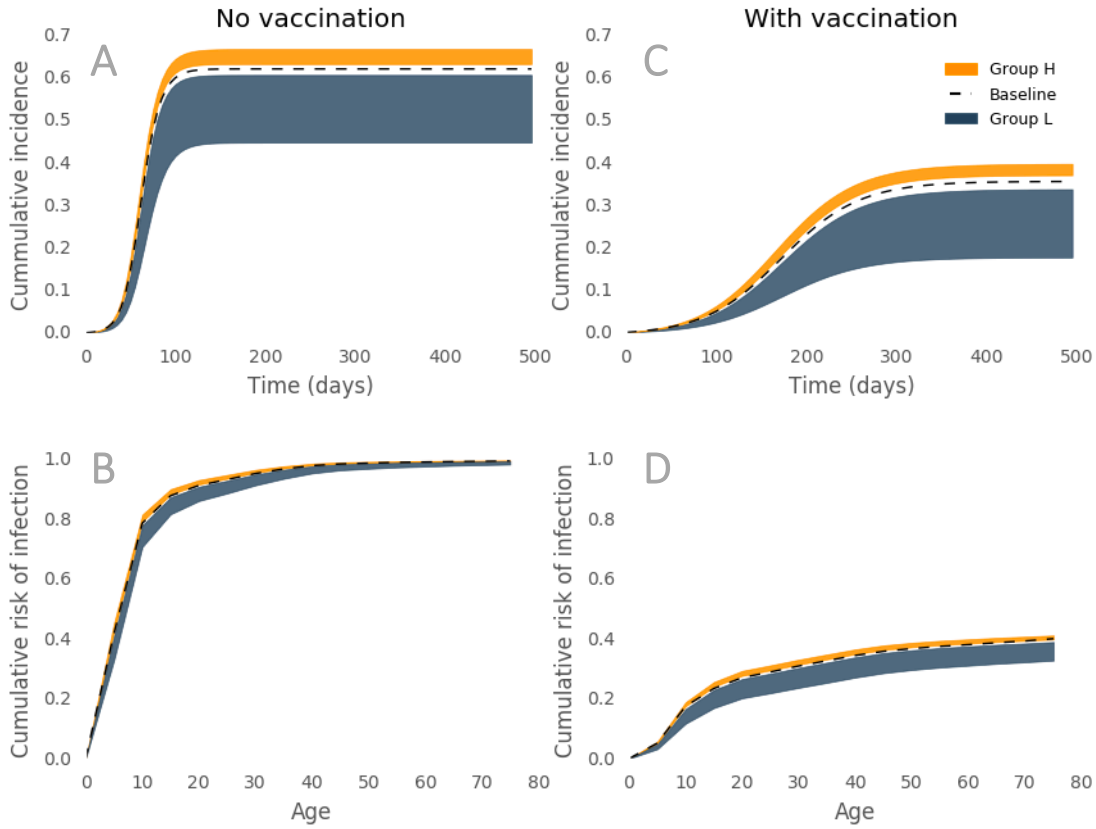


Figure S12 As 10 but with high integration ($\xi = 0.25$)

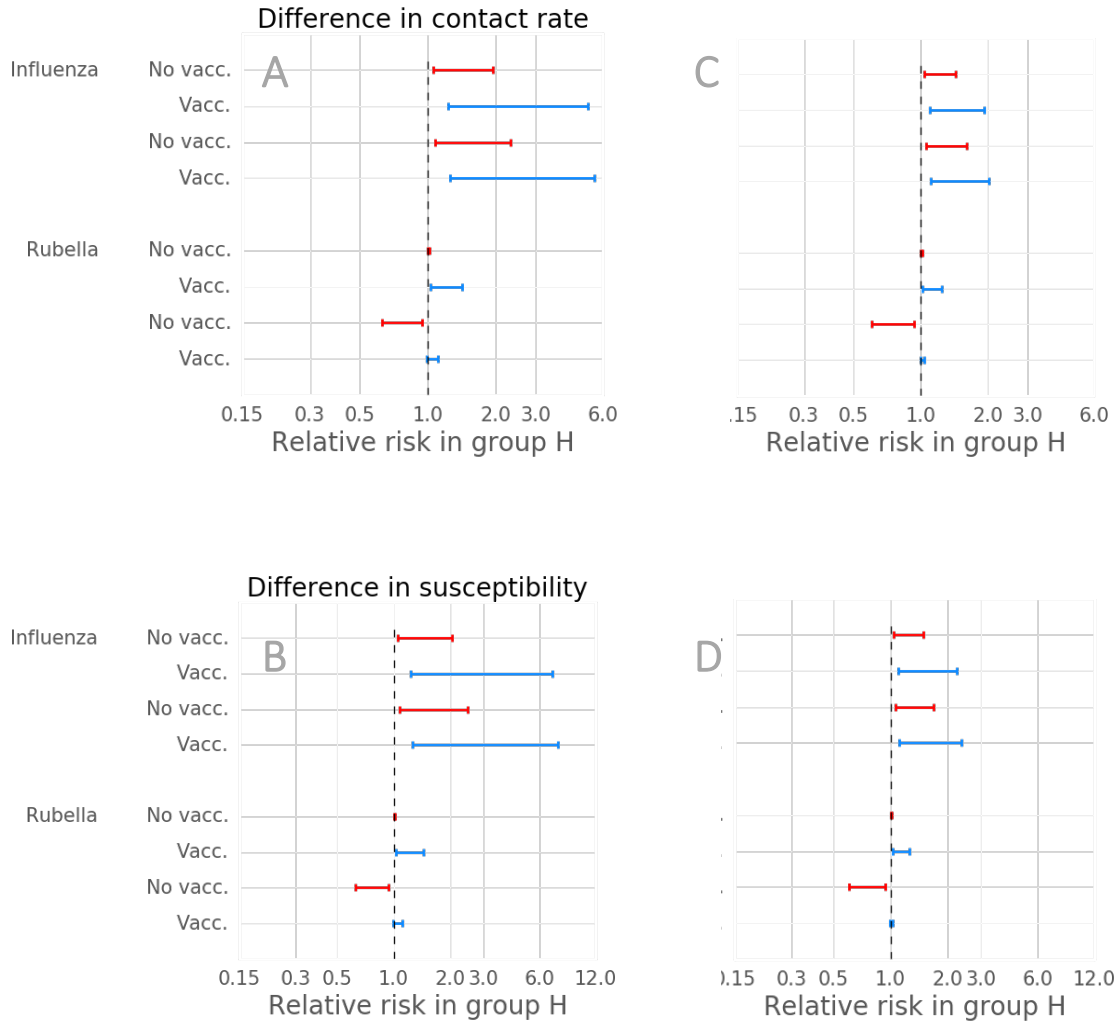


Figure S13 Risk of infection in group H relative to group L in the total population and in risk groups, elderly and women of childbearing age (WCA). Relative risks shown with no vaccination and vaccination at 80% of critical vaccination threshold (37% for influenza and 67% for rubella). Forest plots show ranges of relative risk for a range of ratio of in contact rate in social groups (χ=0.65-0.95) and ratio of susceptibility in social groups (η=0.65-0.95) with integration of ξ = 0.05 (A) and B)) and integration of ξ = 0.25 (C) and D)).

Sensitivity analyses

To test the sensitivity of our results to some of the key assumptions in our model, here we vary the relative size of the social groups and the age-specific mixing.

Size of social groups

The main analysis assumes that the size of each social group is equal (50% of the population each).. Now we assume 80% of the population in the low transmission group, group L, leaving 20% of the population in group H (and vice versa).

Age assortativity in contact matrix

To test the sensitivity of our findings to community structure, we vary relative age-stratified mixing pattern in between the two social groups. First we adjust the 'age-assortativity' (preference in contact with one's own age group over other age groups). We adjusted this using an eigen decomposition method employed by Küchenhoff et al [8], which allows the relative strength of the 'off-diagonal' terms of the mixing matrix to be adjusted with single parameter k. When $k >$ 1 age assortativity decreases (contact between age-groups increases), when $k < 1$ age assortativity increases (contact between age-groups decreases). Transmission matrices for $k = 0.6$ and $k = 1.8$ in Group L are shown in Figure 14.

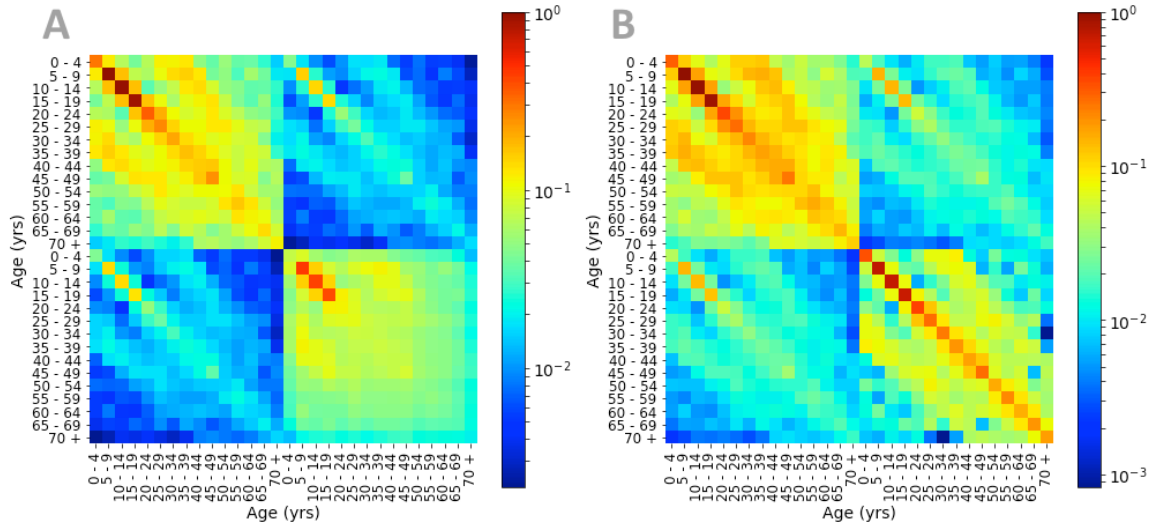


Figure S14 Examples of the full, age and social group structured transmission matrix with A) less age assortativity of Group L (k=1.8) and B) more age assortativity in group L (k=0.6)

We find that changing the age assortativity made little difference to the inequalities we measured, with the exception of rubella in women of childbearing age. In this special case, an increased age assortativity leads to higher rate of contact between children, decreasing the average age at infection in group L. In addition, lower rates of contact between adults and children leads to lower rate of transmission between children and susceptible adults. These factors have the effect of reducing the risk in group L relative to group H with no vaccination. With vaccination, the risk in group L relative to group H reduces as observed in our main analysis.

Conversely, decreasing age assortativity leads to lower rate of contact between children, increasing the average age at infection in group L. Higher rates of contact between adults and children leads to higher rate of transmission between children and susceptible adults. These factors have the effect of increasing the risk in group L relative to group H with no vaccination. With vaccination, the risk in group L relative to group H reduced, as observed in our main analysis.

Although the values for relative risk of rubella infection in women of childbearing age vary notably from the main analysis as assortativity is varied, the qualitative result from our analysis does not change. Namely, ….

Cultural differences in age-specific contact patterns

Here we change the age-specific mixing between social groups by using two distinct empirical social mixing patterns for each social group. Specifically, we assume that group L has age-specific contact rates parameterised with data from the Italian arm of the POLYMOD survey, while group H has age-specific contact rates from the UK arm of the survey. Similar to the results of explicitly changing the age assortativity, we note some changes to relative risks of rubella infection in women of childbearing age. However, these changes did not impact the qualitative result of our analysis (Figure 15 and Figure 16).

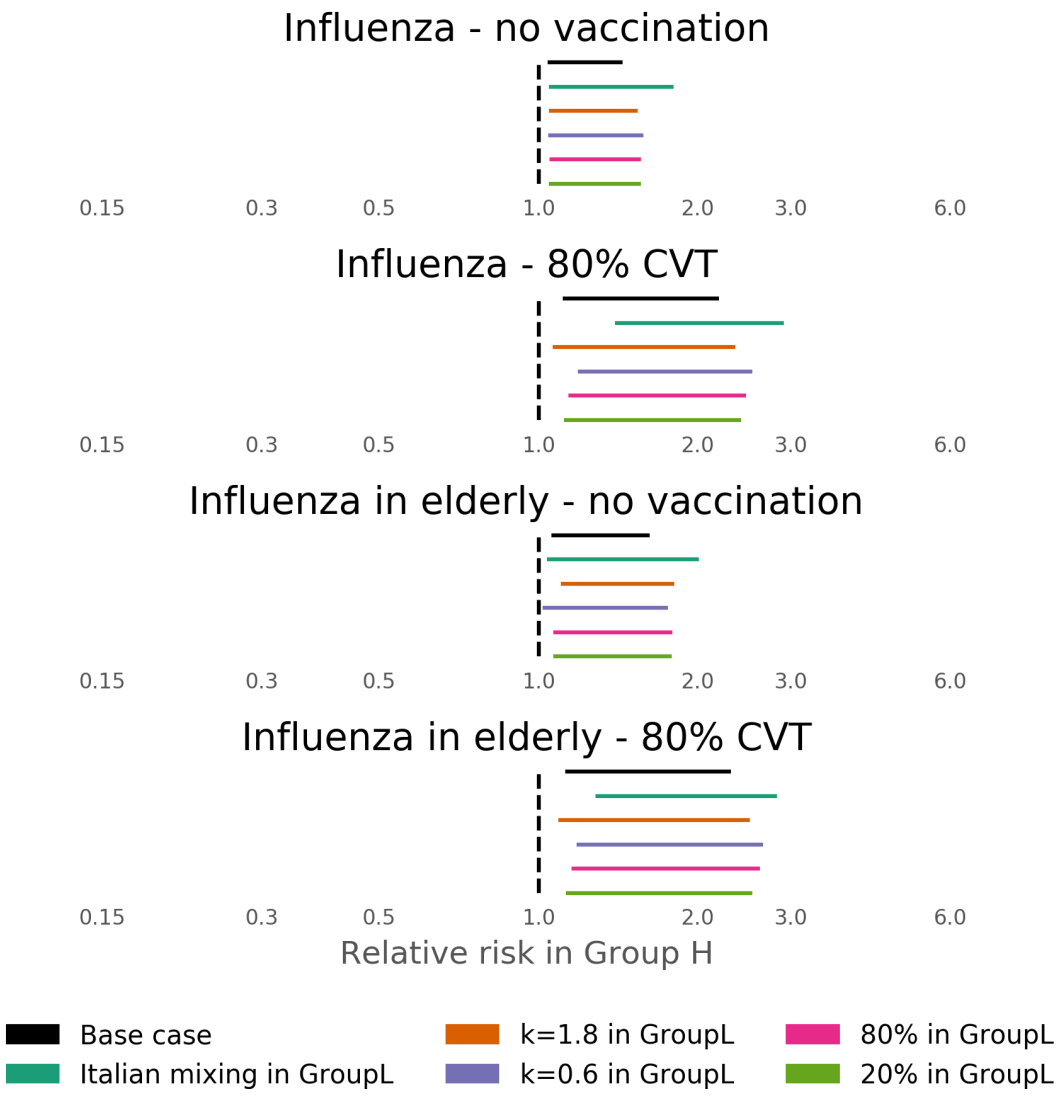


Figure S15 Sensitivity analysis for relative risk in influenza infection due to difference in contact rate (χ=0.65 – 0.95) with integration set as $\xi = 0.15$.

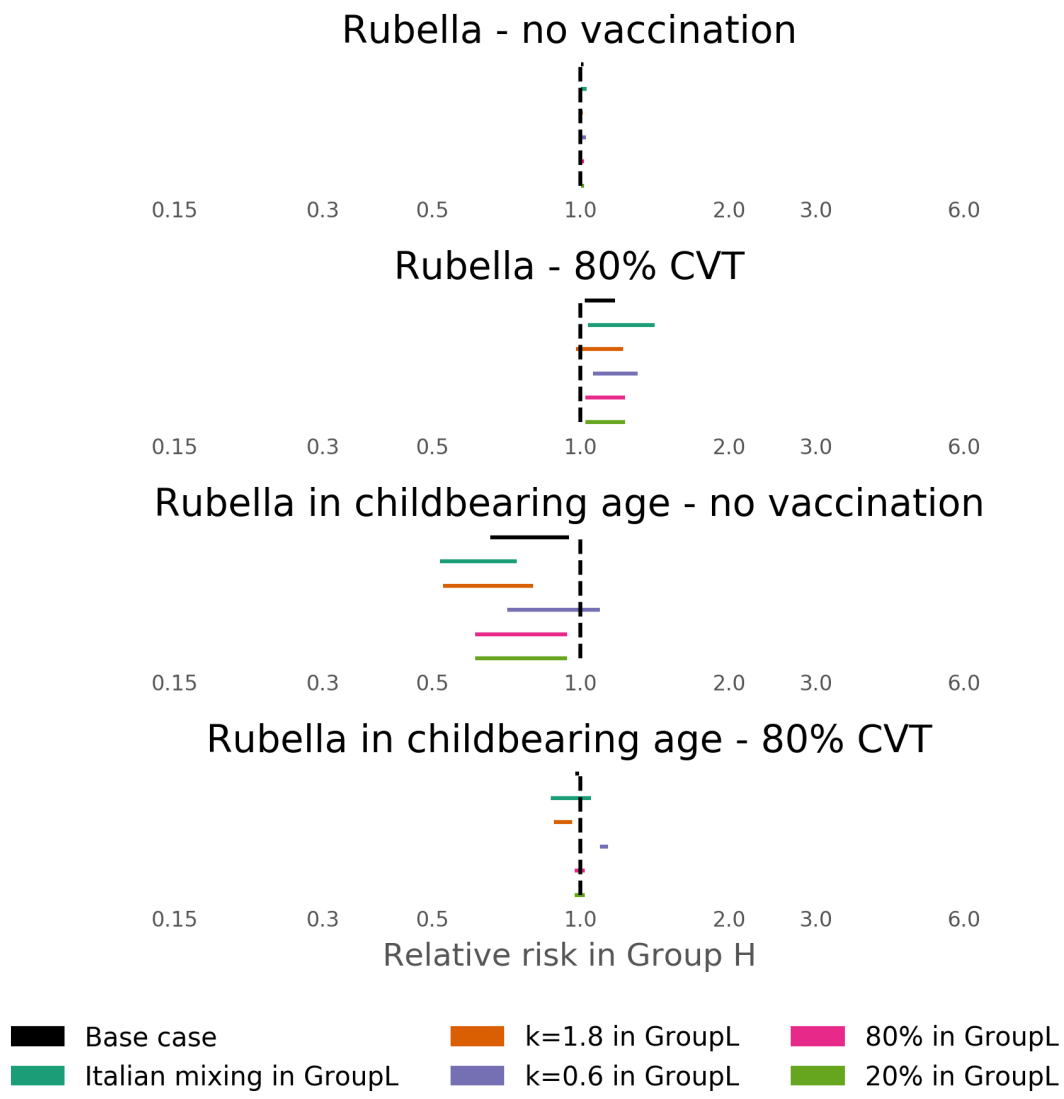


Figure S16 Sensitivity analysis for relative risk in rubella infection due to difference in contact rate (χ=0.65 – 0.95) with integration set as $\xi = 0.15$.

References

1. Mossong J, Hens N, Jit M, Beutels P, Auranen K, Mikolajczyk R, et al. **Social contacts and mixing patterns relevant to the spread of infectious diseases.** *PLoS Med*. 2008, 5:e74.

2. Horby P, Pham QT, Hens N, Nguyen THTHTTY, Le QM, Dang DT, et al. **Social contact patterns in Vietnam and implications for the control of infectious diseases.** *PLoS One*. 2011, 6:e16965.

3. Stein ML, van der Heijden PGM, Buskens V, van Steenbergen JE, Bengtsson L, Koppeschaar CE, et al. **Tracking social contact networks with online respondent-driven detection: who recruits whom?**. *BMC Infect Dis*. 2015, 15:522.

4. Ibuka Y, Ohkusa Y, Sugawara T, Chapman GB, Yamin D, Atkins KE, et al. **Social contacts, vaccination decisions and influenza in Japan.** *J Epidemiol Community Health*. 2016, 70:162–7.

5. Read JM, Lessler J, Riley S, Wang S, Tan LJ, Kwok KO, et al. **Social mixing patterns in rural and urban areas of southern China.** *Proc Biol Sci*. 2014, 281:20140268.

6. Grijalva CG, Goeyvaerts N, Verastegui H, Edwards KM, Gil AI, Lanata CF, et al. **A household-based study of contact networks relevant for the spread of infectious diseases in the highlands of Peru.** *PLoS One*. 2015, 10:e0118457.

7. Béraud G, Kazmercziak S, Beutels P, Levy-Bruhl D, Lenne X, Mielcarek N, et al. **The French Connection: The First Large Population-Based Contact Survey in France Relevant for the Spread of Infectious Diseases.** *PLoS One*. 2015, 10:e0133203.

8. Küchenhoff H, Mwalili SM, Lesaffre E. **A General Method for Dealing with Misclassification in Regression: The Misclassification SIMEX**. *Biometrics*. 2006, 62:85–96.

Appendix B. Supplementary material for Analysis B

Missing data

I evaluated the proportion of reported cases with a missing postcode by date collected, age, gender and case status:

I ran a logistic regression for missing postcode with explanatory variables: Age, date recorded and whether tests were carried out.

Missing postcodes per onset date

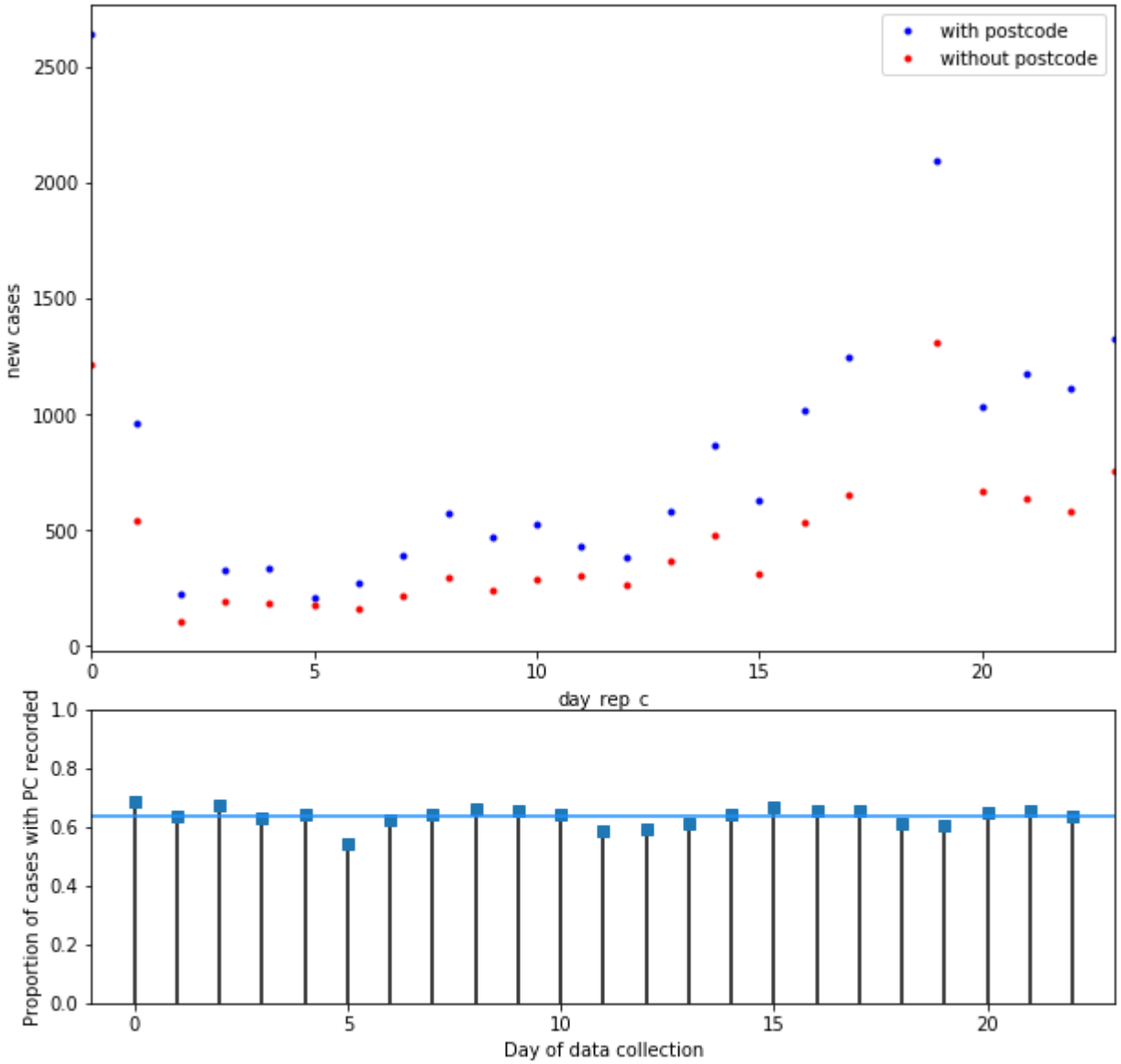


Figure 1 Missing postcodes per onset date

Missing postcodes per age group

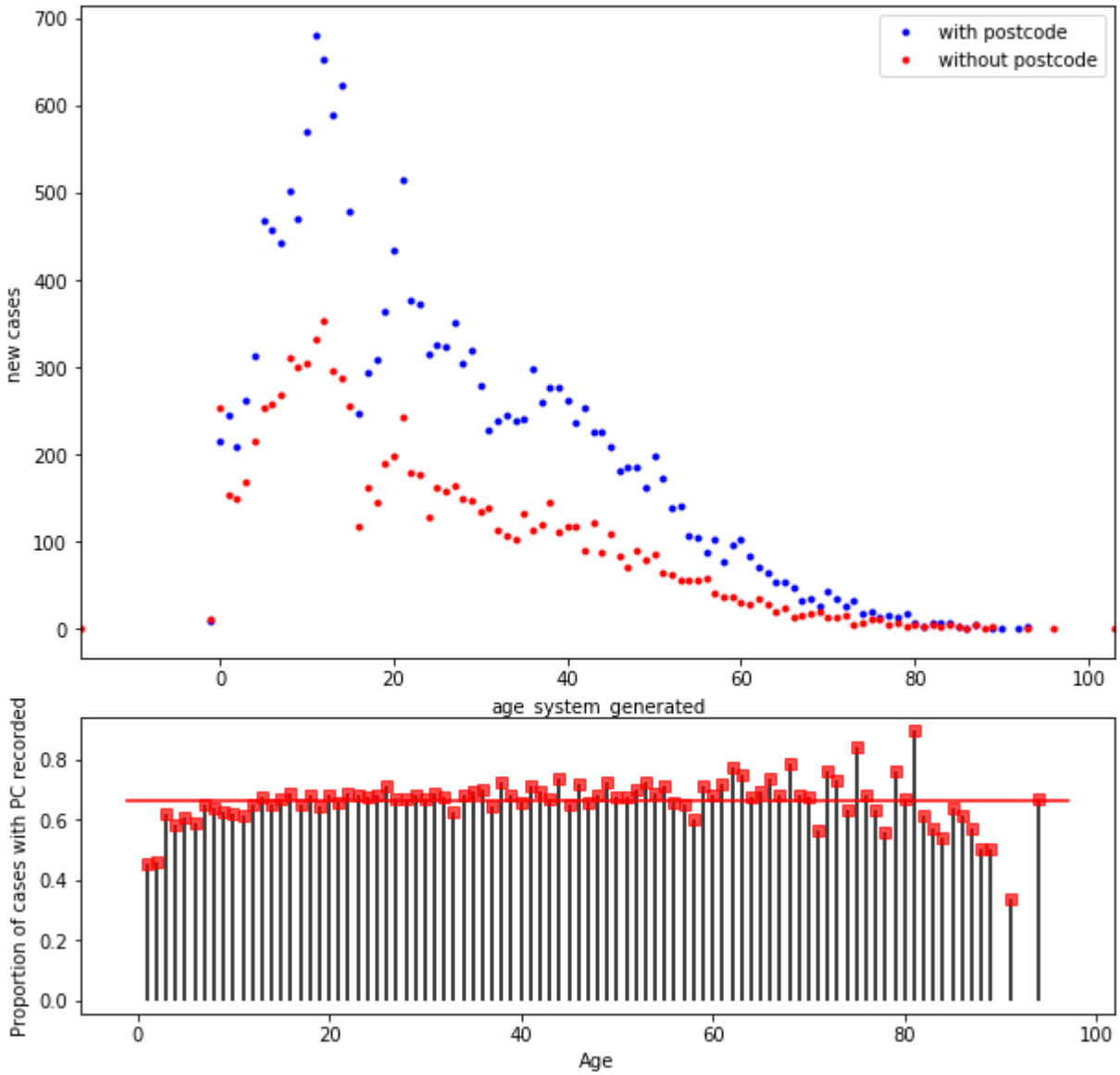


Figure 2 Missing postcodes per age group

Missing postcodes by test status

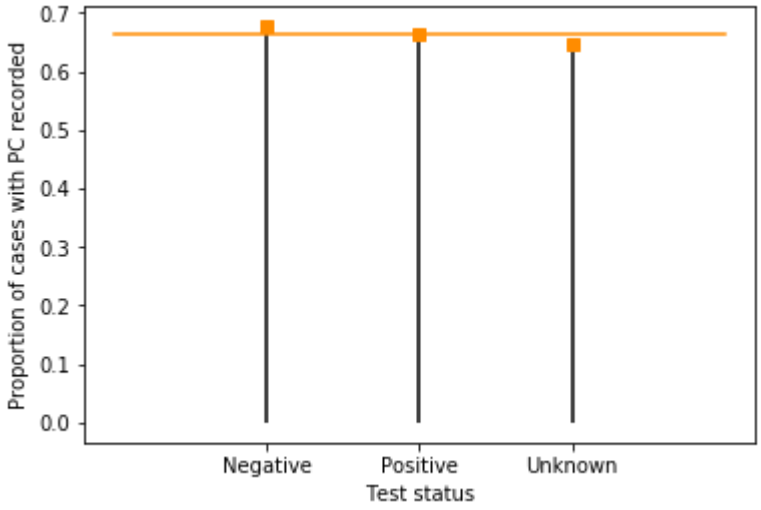


Figure 3 Missing postcodes by test status

Missing postcodes by gender

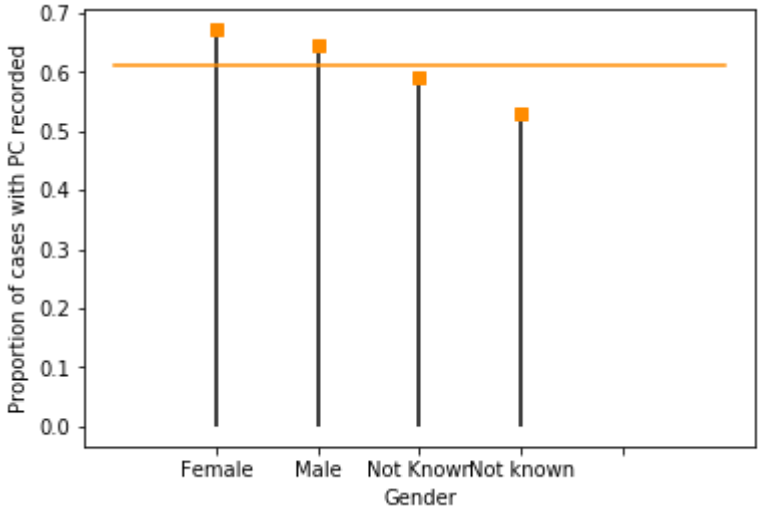


Figure 4 Missing postcodes by gender

Logistic regression for predictors of missing post code:

I performed a logistic regression to identify predictors of a missing postcode in the data.

Table 1 Results from Logistic Regression

Dep. Variable:	PC_pres	**No. Observations:**	28428
Model:	Logit	**Df Residuals:**	28424
Method:	MLE	**Df Model:**	3
Date:	Wed, 08 Nov 2017	**Pseudo R-squ.:**	0.004056
Time:	17:21:05	**Log-Likelihood:**	-18108.
converged:	True	**LL-Null:**	-18182.
		LLR p-value:	9.091e-32

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	0.4115	0.037	10.985	0.000	0.338	0.485
age_system_generated	0.0062	0.001	8.315	0.000	0.005	0.008
day_rep_c	-0.0010	0.002	-0.619	0.536	-0.004	0.002
tested	0.2105	0.026	8.122	0.000	0.160	0.261

2 Percentage of cases tested and tests positive by ethnic group

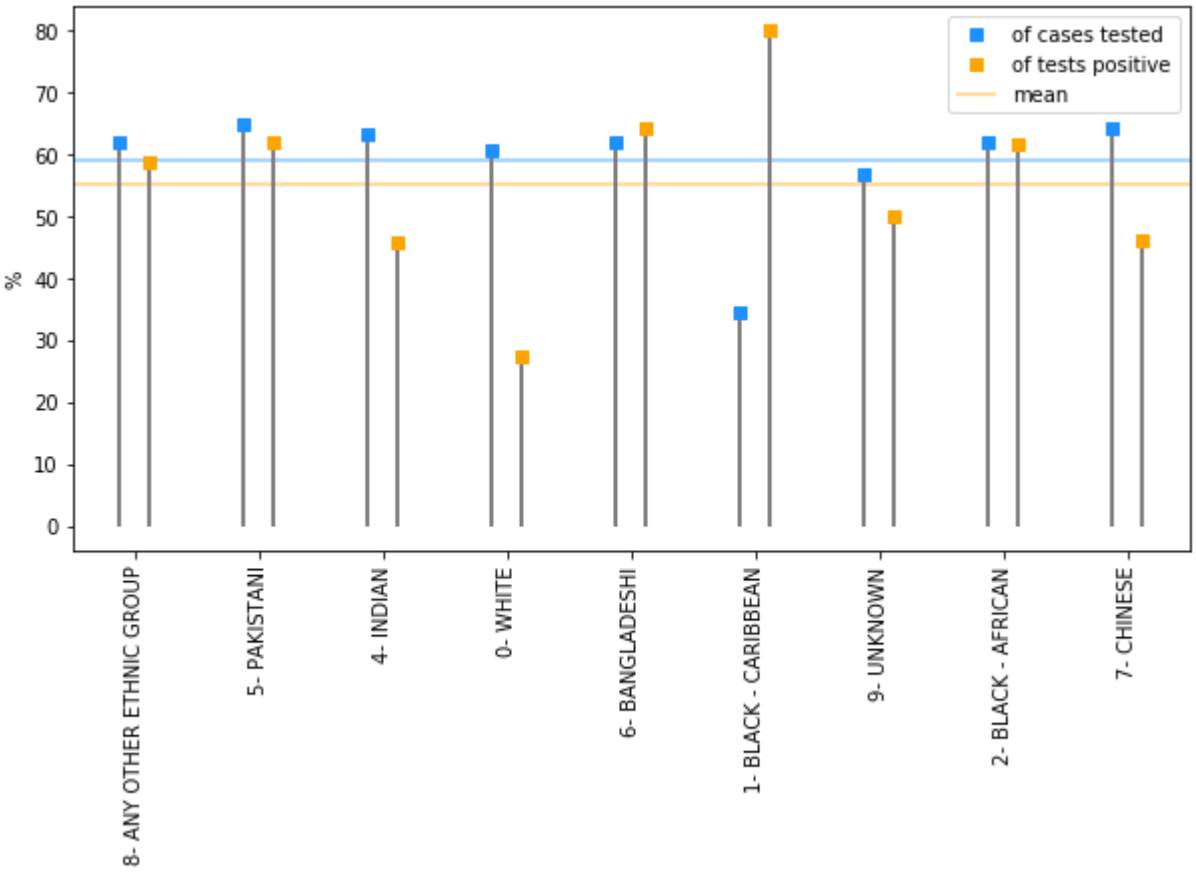


Figure 5 Percentage of cases tested and tests positive by ethnic group

- The proportion of tests that were positive is much lower in the white population. This could mean that there was an inflated reporting rate in this population.

- Noticeably, the test rate of cases was consistent for most ethnicities (around 60-65%) but this drops to less than 40% in black caribbeans. There were very small number of cases though - so may not be significant.

Additional results:

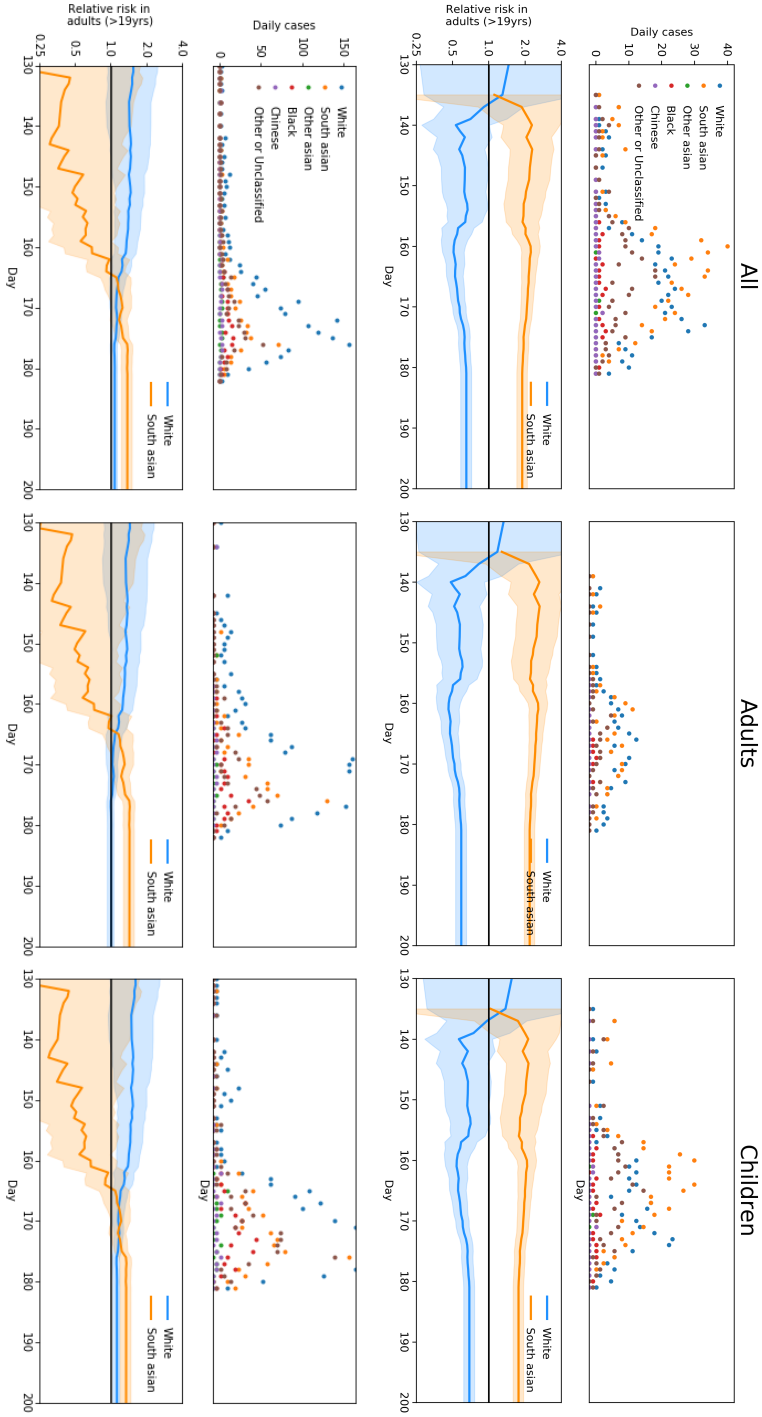


Figure 6 Ethnic breakdown of cases. i) Daily incidence in each ethnic group identified by Onomap, ii) Relative risk in White (Blue) and South Asian (Orange) populations for A) Birmingham , B) London, C) Adults in Birmingham (>19), D) Adults in London (>19), E) Children in Birmingham (<=19) and F) Children in London (<=19).

Appendix C. Suplementary material for Analysis C

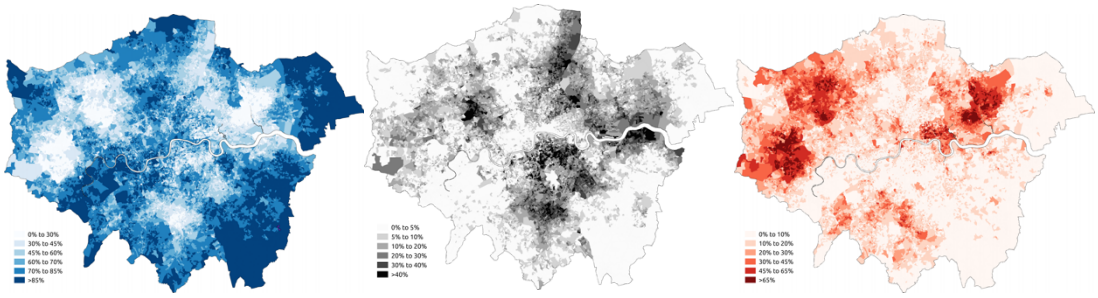


Figure 1 Choropleths with distribution of A) White, B) Black and C) Asian population in London.

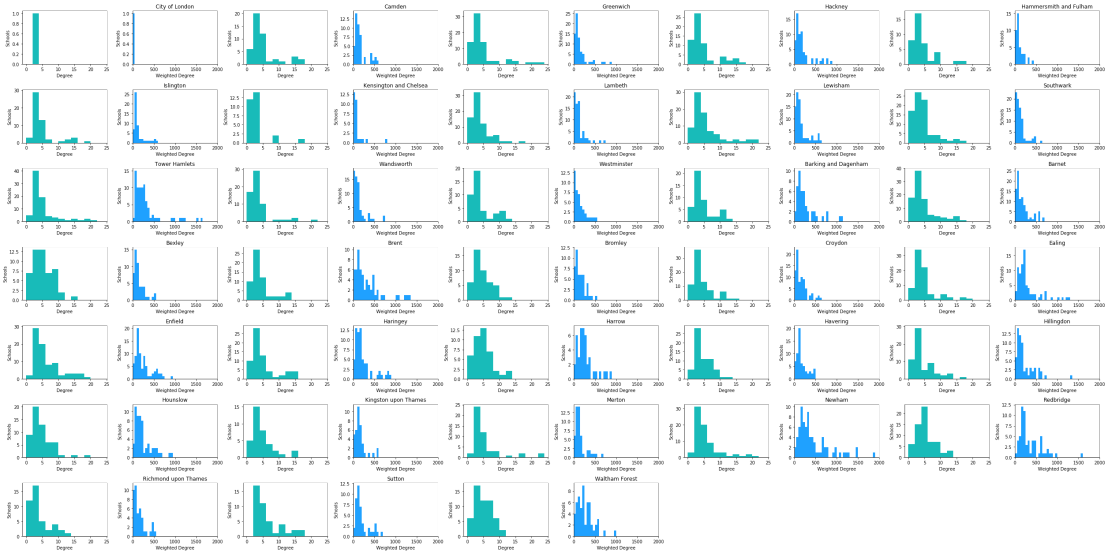


Figure 2 Degree and weighted degree histograms per London Borough

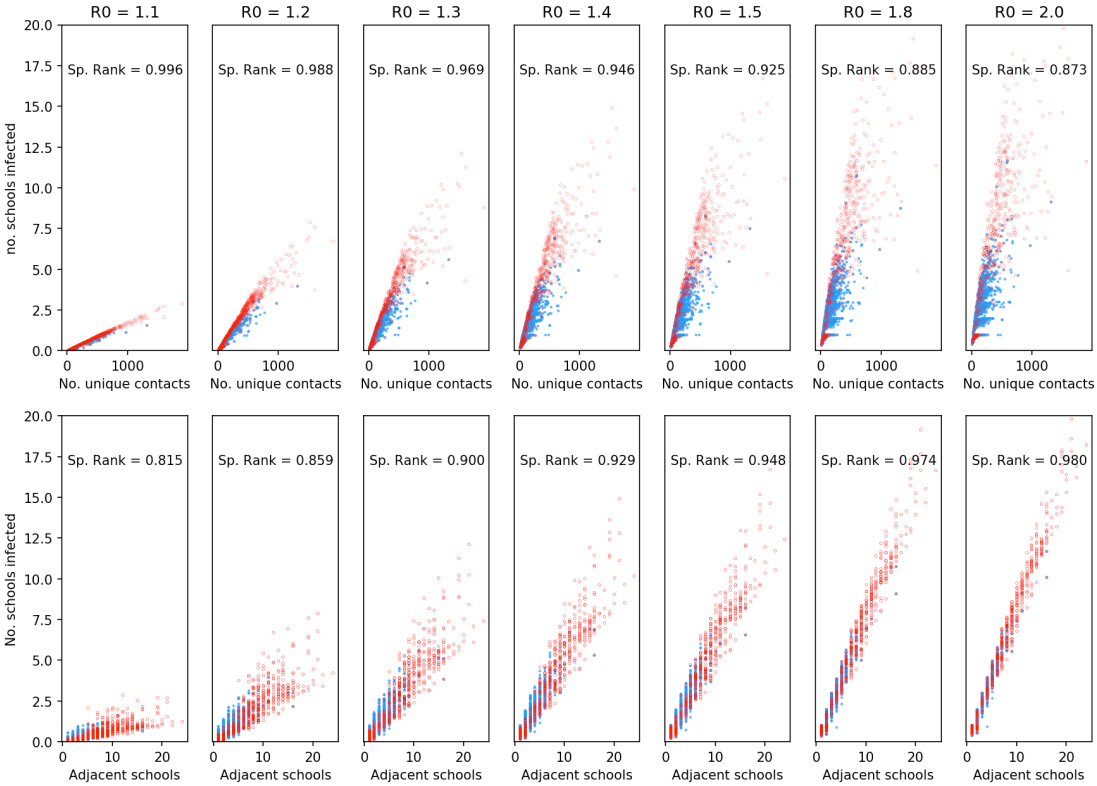


Figure 3 Scatter plots of expected number of infected schools and number of unique contact pairs (weighted degree of contact the school network) and number of adjacent schools (degree of the school contact network) for primary (blue) and secondary (red) schools in London, for different values of R_0.

	Mean no. seeded						Relative difference					
	R0 = 1.1	*R0 = 1.2*	*R0 = 1.3*	*R0 = 1.4*	*R0 = 1.5*	*R0 = 1.8*	*R0 = 2.0*	*R0 = 1.1*	*R0 = 1.2*	*R0 = 1.3*	*R0 = 1.4*	*R0 = 1.5*	*R0 = 1.8*	*R0 = 2.0*
White British	0.97	2.96	4.79	6.20	7.25	9.09	9.72	0.87	0.90	0.92	0.93	0.94	0.95	0.95
White; Irish	1.04	3.14	5.03	6.48	7.55	9.39	10.01	0.94	0.95	0.96	0.97	0.97	0.98	0.98
White Traveller	1.13	3.38	5.36	6.87	7.97	9.84	10.47	1.02	1.02	1.03	1.03	1.03	1.03	1.03
Other White	1.08	3.26	5.20	6.68	7.77	9.62	10.23	0.98	0.99	1.00	1.00	1.00	1.00	1.01
White and Black Caribbean	1.05	3.18	5.12	6.61	7.72	9.63	10.27	0.95	0.97	0.98	0.99	1.00	1.01	1.01
White and Black African	1.08	3.26	5.20	6.69	7.78	9.65	10.27	0.98	0.99	1.00	1.00	1.00	1.01	1.01
White and Asian	1.08	3.24	5.15	6.61	7.68	9.51	10.12	0.98	0.98	0.99	0.99	0.99	0.99	0.99
Other Mixed	1.07	3.23	5.16	6.64	7.74	9.61	10.24	0.97	0.98	0.99	0.99	1.00	1.00	1.01
Indian	1.44	4.03	6.06	7.49	8.49	10.08	10.59	1.31	1.22	1.16	1.12	1.09	1.05	1.04
Pakistani	1.51	4.15	6.21	7.65	8.65	10.25	10.75	1.37	1.26	1.19	1.14	1.12	1.07	1.06
Bangladeshi	1.63	4.67	7.12	8.83	10.00	11.78	12.31	1.47	1.42	1.36	1.32	1.29	1.23	1.21
Chinese	1.03	3.13	5.04	6.51	7.60	9.46	10.09	0.93	0.95	0.97	0.97	0.98	0.99	0.99
Other Asian	1.25	3.62	5.62	7.10	8.17	9.94	10.53	1.13	1.10	1.08	1.06	1.05	1.04	1.03
Black African	1.11	3.32	5.29	6.79	7.90	9.79	10.42	1.00	1.01	1.01	1.02	1.02	1.02	1.02
Black Caribbean	1.11	3.34	5.33	6.86	7.98	9.90	10.54	1.01	1.01	1.02	1.03	1.03	1.03	1.04
Other Black	1.11	3.33	5.30	6.81	7.93	9.83	10.46	1.00	1.01	1.02	1.02	1.02	1.03	1.03
Arab	1.14	3.38	5.32	6.76	7.81	9.56	10.13	1.04	1.03	1.02	1.01	1.01	1.00	1.00
Other Ethnic Group	1.16	3.46	5.47	6.98	8.09	9.95	10.57	1.05	1.05	1.05	1.04	1.04	1.04	1.04

	Proportion infected						Relative difference					
	R0 = 1.1	*R0 = 1.2*	*R0 = 1.3*	*R0 = 1.4*	*R0 = 1.5*	*R0 = 1.8*	*R0 = 2.0*	*R0 = 1.1*	*R0 = 1.2*	*R0 = 1.3*	*R0 = 1.4*	*R0 = 1.5*	*R0 = 1.8*	*R0 = 2.0*
White British	0.35	0.21	0.16	0.11	0.08	0.08	0.06	1.09	1.10	1.11	1.11	1.11	1.11	1.11
White; Irish	0.34	0.21	0.15	0.11	0.08	0.07	0.06	1.06	1.07	1.07	1.07	1.07	1.07	1.07
White Traveller	0.31	0.18	0.14	0.09	0.07	0.06	0.05	0.96	0.95	0.94	0.94	0.94	0.94	0.94
Other White	0.34	0.20	0.15	0.11	0.08	0.07	0.06	1.04	1.05	1.05	1.05	1.06	1.06	1.06
White and Black Caribbean	0.34	0.20	0.15	0.11	0.08	0.07	0.06	1.04	1.04	1.04	1.04	1.04	1.04	1.04
White and Black African	0.33	0.20	0.15	0.10	0.08	0.07	0.06	1.02	1.02	1.02	1.02	1.02	1.02	1.02
White and Asian	0.33	0.20	0.15	0.10	0.08	0.07	0.06	1.03	1.03	1.03	1.04	1.04	1.04	1.04
Other Mixed	0.33	0.20	0.15	0.11	0.08	0.07	0.06	1.04	1.04	1.04	1.04	1.04	1.04	1.04
Indian	0.24	0.14	0.10	0.07	0.05	0.05	0.04	0.74	0.71	0.71	0.70	0.70	0.70	0.70
Pakistani	0.23	0.13	0.10	0.07	0.05	0.04	0.04	0.70	0.67	0.66	0.66	0.65	0.65	0.65
Bangladeshi	0.23	0.13	0.10	0.07	0.05	0.04	0.04	0.71	0.68	0.67	0.66	0.66	0.65	0.65
Chinese	0.35	0.21	0.16	0.11	0.08	0.08	0.06	1.09	1.10	1.11	1.11	1.11	1.11	1.11
Other Asian	0.28	0.17	0.12	0.09	0.06	0.06	0.05	0.88	0.86	0.85	0.85	0.85	0.85	0.85
Black African	0.32	0.19	0.14	0.10	0.08	0.07	0.06	1.00	0.99	0.99	0.99	0.99	0.99	0.99
Black Caribbean	0.32	0.19	0.14	0.10	0.07	0.07	0.05	0.98	0.97	0.97	0.97	0.97	0.97	0.97
Other Black	0.32	0.19	0.14	0.10	0.08	0.07	0.06	1.00	0.99	0.99	0.99	0.99	0.99	0.99
Arab	0.32	0.19	0.14	0.10	0.08	0.07	0.06	0.99	0.99	0.99	0.99	0.99	0.98	0.98
Other Ethnic Group	0.31	0.18	0.14	0.10	0.07	0.06	0.05	0.96	0.95	0.95	0.94	0.94	0.94	0.94

Table 1 Mean Expected number of schools infected and proportion of school infected before seeding an outbreak in an adjacent school for each ethnic group

Appendix D. Suplementary material for Analysis D (part 1)

provincie	Denomination	All schools	In Consesnsus community	%
Gelderland	Protestants-Christelijk	356	6	1.7%
	Reformatorisch	52	42	80.8%
Noord-Brabant	Reformatorisch	4	4	100.0%
Noord-Holland	Reformatorisch	2	2	100.0%
Overijssel	Reformatorisch	22	7	31.8%
Utrecht	Protestants-Christelijk	207	4	1.9%
	Reformatorisch	16	16	100.0%
Zeeland	Algemeen bijzonder	22	19	86.4%
	Antroposofisch	1	1	100.0%
	Gereformeerd vrijgemaakt	3	3	100.0%
	Openbaar	96	79	82.3%
	Overige	1	1	100.0%
	Protestants-Christelijk	63	56	88.9%
	Reformatorisch	36	36	100.0%
	Rooms-Katholiek	48	45	93.8%
	Samenwerking PC, RK	10	10	100.0%
Zuid-Holland	Algemeen bijzonder	169	1	0.6%
	Protestants-Christelijk	524	18	3.4%
	Reformatorisch	65	59	90.8%

Table 1 Breakdown, by Province and Denomination, of community with majority of Dutch Reformed schools in community with highest modularity.

provincie	Denomination	All schools	In Consesnsus community	%
Gelderland	Protestants-Christelijk	356	3	0.8%
	Reformatorisch	52	38	73.1%
Noord-Holland	Reformatorisch	2	1	50.0%
Utrecht	Protestants-Christelijk	207	2	1.0%
	Reformatorisch	16	14	87.5%
Zuid-Holland	Protestants-Christelijk	524	8	1.5%
	Reformatorisch	65	54	83.1%

Table 2 Breakdown, by Province and Denomination, of the group of nodes which were partitioned into the same community in every initial partition, with the highest proportion of Dutch Reformed schools.

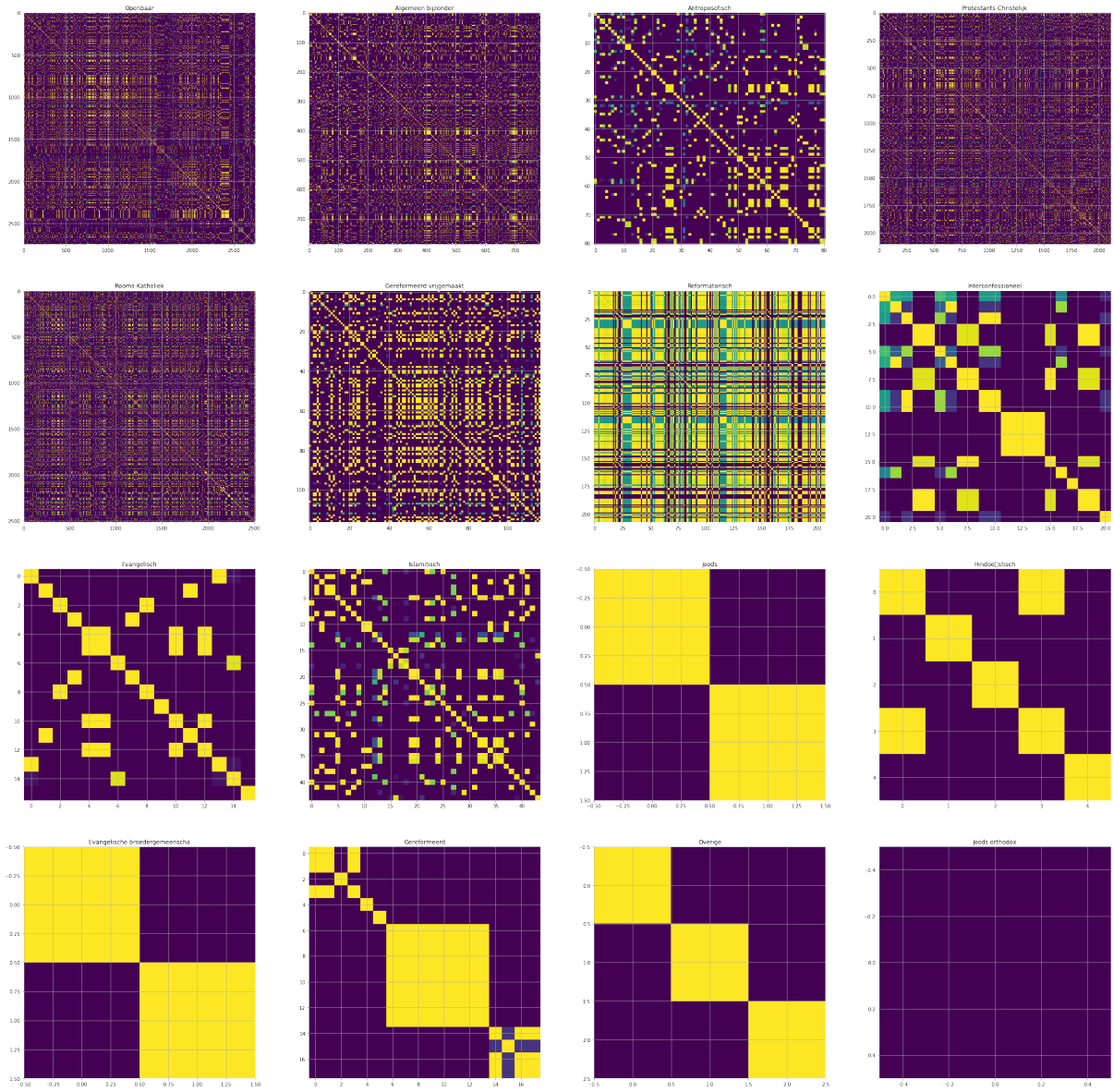


Figure 1 Subset of the similarity matrix for schools affiliated with each denomination (schools affiliated with multiple denominations excluded, denominations with only one school excluded). Entries in the matrix show as purple low to yellow high proportion of partitions with schools in the same community.

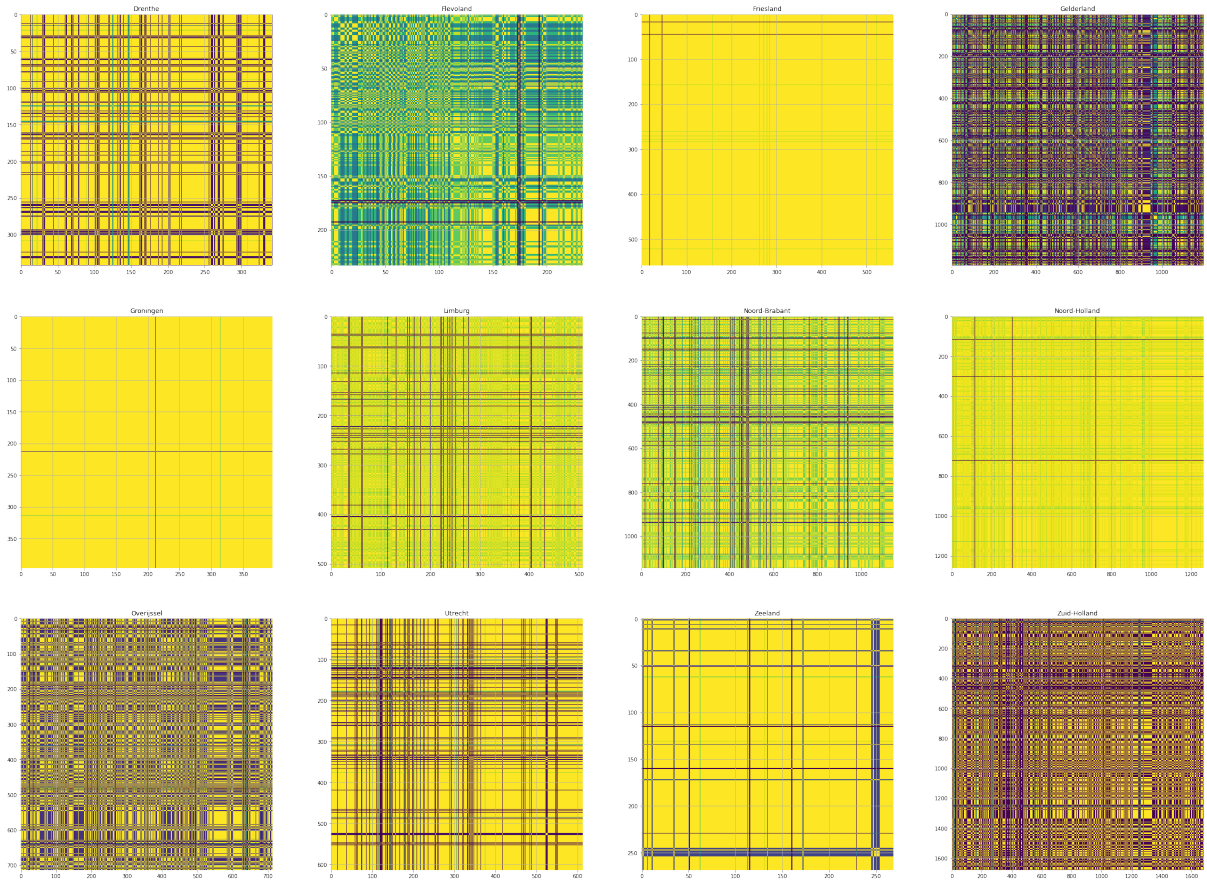


Figure 2 Subset of the similarity matrix for schools in each Province. Entries in the matrix show as purple low to yellow high proportion of partitions with schools in the same community.

Appendix E. Suplementary material for Analysis D (part 2)

Alternative Model 2: Spatial interaction between schools

To evaluate the importance of the specific connections between schools in the network, I developed an Alternative Model of interaction between schools based on their geographic proximity to each other. The spatial interaction model was designed to have a broadly equivalent spatial distribution of contact between schools but with otherwise naïve interaction (i.e. no preference for contact within particular denominations or between tiers of education etc.). to do this I made attempt to match the distribution of the *distance between schools connected by contact pairs*.

The spatial interaction parameter used in the analysis was 0.65, which was aimed at matching the peak of the distribution of contact described in the data.

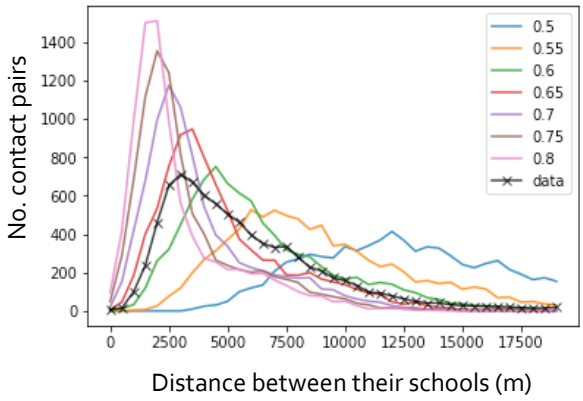


Figure 1 The distribution of distance between schools connected by contact pairs. The coloured lines show interaction based on spatial interaction model with parameters between 0.5 and 0.8. black line with cross markers shows the distribution in the household links from the data.

Risk by schools calculations

The binary outbreak networks were used to identify schools, which would be infected by an outbreak initiated in each school in the network. This was achieved using the following approach:

The binary outbreak network is a directional graph with edges weighted either 1 or 0. 1 means transmission occurs between the schools in the direction of the edge.

The model can calculate both risk posed by each school (the number of schools and children eventually infected by an outbreak initiated in the school) and the risk posed to a school (the number of schools or children who could initiate an outbreak which would eventually infect the school). Both of these are found by following chains of edges away from the school of interest. The difference between extracting these from a binary outbreak network is the "direction of transmission" of the edges followed.

To calculated the risk posed by a school, the schools eventually infected are those connected by chains of out-edges (schools infected by the previous generation of the outbreak) on the binary outbreak network. Figure A5.2 shows schools in progressive generations of an outbreak on an example network. The school of interest (i.e. the school posing the risk) is coloured red. In each network shows a different generation of the outbreak, where newly infected schools are shown in green and previous generations are shown in blue. There are 4 generations of transmission after the initial school infected and10 schools are eventually infected by the outbreak.

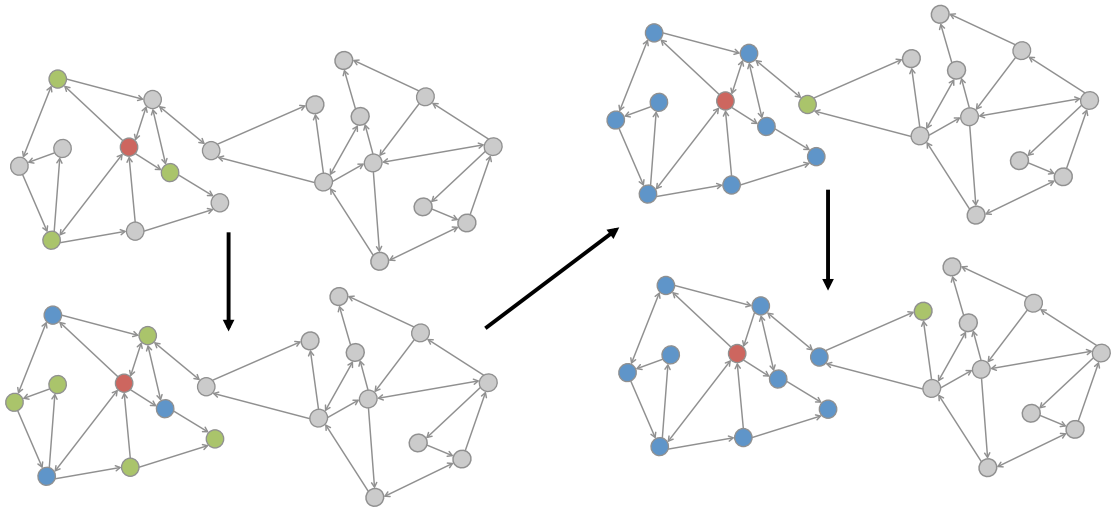


Figure 2 Schematic of a network showing chains of out-edges to form the out component of a node as a method for finding the schools at risk of infection from an outbreak initiated in a particular school. The initial school is shown in red, each new generation of schools connected by out-edges is shown in green, previous generations are shown in blue. Black arrows indicate the direction of successive generations

To calculate the risk posed to a school, the schools who could seed an outbreak that would eventually infect the school are those connected by chains of in-edges (schools who would infect each generation of the outbreak) on the binary outbreak network. Figure A5.2 shows schools in regressive generations of an outbreak on an example network. The school of interest (i.e. the school posing the risk) is coloured red. Each network shows a different generation of the outbreak, where newly infected schools are shown in green and following generations are shown in blue. There are a maximum of 7 generations of transmission after with the school at risk is infected. There are schools are eventually infected. Outbreaks initiated in 13 schools would eventually infect the school.

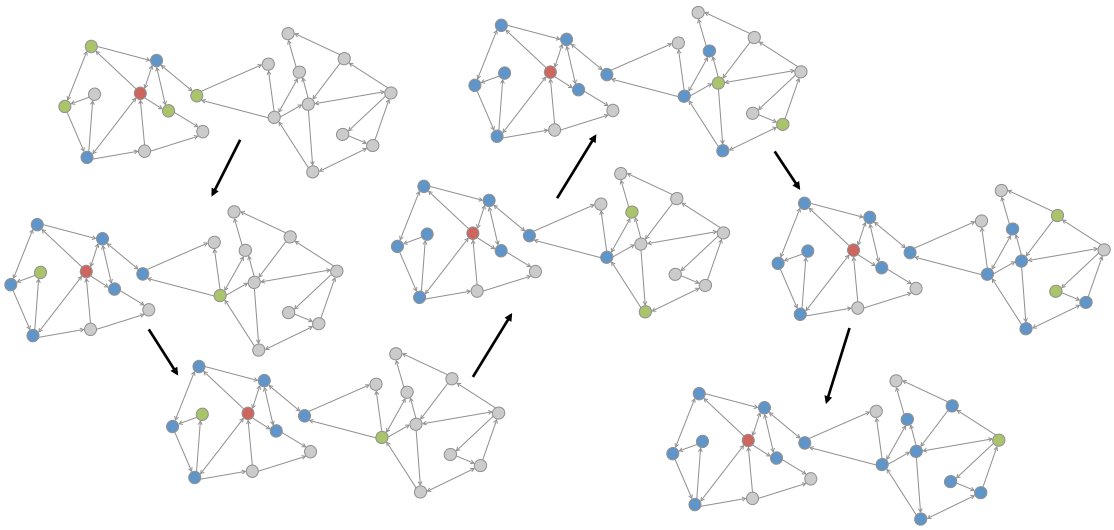


Figure 3 Schematic of a network showing chains of in-edges to form the in component of a node as a method for finding the schools from an outbreak initiated in a particular school. The school at risk is shown in red, each previous possible generation of schools connected by in-edges is shown in green, onward generations are shown in blue. Black arrows indicate the direction of regressive generations

Appendix F. LSHTM ethics approval

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

Observational / Interventions Research Ethics Committee

Mr James Munday
LSHTM

15 March 2019

Dear James

Study Title: National schools data from the Netherlands

LSHTM Ethics Ref: 16028 - 1

Thank you for your application for the above amendment to the existing ethically approved study and submitting revised documentation The amendment application has been considered by the Observational Committee via Chair's Action.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above amendment to research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval for the amendment having been received, where relevant.

Approved documents

The final list of documents reviewed and approved is as follows:

Document Type	File Name	Date	Version
Other	Protocol_ammendment_1	04/03/2019	2

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using the End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk.

Further information is available at: www.lshtm.ac.uk/ethics.

Yours sincerely,

[redacted]

Professor John DH Porter
Chair

ethics@lshtm.ac.uk
http://www.lshtm.ac.uk/ethics/

London School of Hygiene & Tropical Medicine
Keppel Street, London WC1E 7HT
United Kingdom
Switchboard: +44 (0)20 7636 8636

www.lshtm.ac.uk

Observational / Interventions Research Ethics Committee

Mr James Munday
LSHTM

2 August 2018

Dear James,

Study Title: Analysis of inequalities in exposure to Influenza A H1N1 during 2009 pandemic

LSHTM Ethics Ref: 14559

Thank you for responding to the Observational Committee's request for further information on the above research and submitting revised documentation.

The further information has been considered on behalf of the Committee by the Chair.

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form, protocol and supporting documentation as revised, subject to the conditions specified below.

Conditions of the favourable opinion

Approval is dependent on local ethical approval having been received, where relevant.

Approved documents

The final list of documents reviewed and approved by the Committee is as follows:

Document Type	File Name	Date	Version
Investigator CV	JamesMunday CV	17/11/2017	1
Investigator CV	CV AJ van Hoek	15/12/2017	1
Protocol / Proposal	Protocol	17/05/2018	1
Covering Letter	EthicsCoverLetter	05/07/2018	1

After ethical review

The Chief Investigator (CI) or delegate is responsible for informing the ethics committee of any subsequent changes to the application. These must be submitted to the Committee for review using an Amendment form. Amendments must not be initiated before receipt of written favourable opinion from the committee.

The CI or delegate is also required to notify the ethics committee of any protocol violations and/or Suspected Unexpected Serious Adverse Reactions (SUSARs) which occur during the project by submitting a Serious Adverse Event form.

An annual report should be submitted to the committee using an Annual Report form on the anniversary of the approval of the study during the lifetime of the study.

At the end of the study, the CI or delegate must notify the committee using an End of Study form.

All aforementioned forms are available on the ethics online applications website and can only be submitted to the committee via the website at: http://leo.lshtm.ac.uk

Additional information is available at: www.lshtm.ac.uk/ethics

Yours sincerely,

[redacted]

Professor John DH Porter
Chair

ethics@lshtm.ac.uk
http://www.lshtm.ac.uk/ethics/

Appendix G. License for re-publication of BMC Medicine paper

Extract from the BMC website regarding reproduction of published works:

https://www.biomedcentral.com/getpublished/copyright-and-license

Copyright and License

- Copyright on any open access article in a journal published by BMC is retained by the author(s).
- Authors grant BMC a license to publish the article and identify itself as the original publisher.
- Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified.
- The Creative Commons Attribution License 4.0 formalizes these and other terms and conditions of publishing articles.
- In accordance with our Open Data policy, the Creative Commons CC0 1.0 Public Domain Dedication waiver applies to all published data in BMC open access articles.

Where an author is prevented from being the copyright holder (for instance in the case of US government employees or those of Commonwealth governments), minor variations may be required. In such cases the copyright line and license statement in individual articles will be adjusted, for example to state '© 2016 Crown copyright'. Authors requiring a variation of this type should inform BMC during or immediately after submission of their article. Changes to the copyright line cannot be made after publication of an article.